Insurance Cost Estimation by Supervised Regression Approach

Tonni Das Jui

Abstract

Since its emergence, health insurance has increased in popularity in both developed and developing countries. Underdeveloped countries are now working toward developing a healthcare system to assist people in avoiding bankruptcy due to medical expenses. Individuals who have health insurance have the advantage of paying a lower amount for medical care than the bill, which helps them deal with unexpectedly high medical bills. Government or private entities offer health care packages. For several organizations, such as hospitals, an accurate estimation of health costs is important. It necessitates the prediction of insurance costs based on an analytical dataset of important individual variables such as age, average monthly medical costs, and so on. In this paper, we investigate the dataset from Brett Lantz's book Machine Learning with R which was simulated using demographic statistics in the United States, and we compare the efficiency of various models in predicting insurance costs on this dataset. To predict the cost, we employed Linear Regression, Ridge Regression, Support Vector Regression, Regression with Neural Network, and Random Forest Regression, and we faced overfitting issues by running several experiments with tuning the hyper-parameters of the best performing model.

Keywords— Insurance cost estimation, EDA, Linear Regression, Ridge regression, Neural Network, Support Vector regression, Random forest, Polynomial feature transformation, Hyperparameter Tuning, Cross validation

1 Introduction

Rising medical costs and treatment for illnesses have made arranging funds during a medical emergency a daunting task. One clinical research study published in the American journal of medicine around 2007 claimed that 62.1% of bankruptcies were caused by medical issues [1]. Basic health insurance policy can be two types, one is public that is offered by the state and another is private that is offered by various insurance companies. These companies determine multiple insurance plans well suited for individuals with beneficial coverage. These insurance plans ensure the financial security, domiciliary treatment, free health check-up, tax benefits, coverage against critical illness, etc.

The primary responsibility of a health insurance provider is to estimate an individual's insurance expense allocation. To predict the insurance cost, it requires to collect years of data set an individual spends on medical care. Age, gender, eating habits, daily budget, income range, BMI, dependents, smoking habits, living location, and other factors may have an impact on a person's medical costs. The dataset that we are analyzing contains the majority of the above-mentioned simple features and complex features.

The objective of this project is to predict health insurance cost for an individual exploring the collected dataset. The key motivation for this project is to help health insurance policymakers to allocate health insurance boundaries more suited to specific individuals' lifestyles for specific geographical locations. This will assist them in investigating individual medical insurance costs vs. the appropriate insurance plan model. Our goal is to consider these factors and estimate the tentative medical insurance costs of an individual.

The estimation of approximate medical costs on a monthly or annual basis is a regression problem. Predicting a real or continuous goal output value based on several simple and complex features necessitates the use of regression models. In this project, we study the nature of predicted and actual values of medical cost and judge the performance of linear and ridge regression, support vector machine, regression with neural

network, and polynomial regression. We also study the effect of tuning hyperparameter on the random forest model for the dataset.

2 Literature Review

One of the world's most important concerns is the rising cost of health care, and many people have invested in forecasting health costs to avoid bankruptcy due to medical costs. In this regard, Various data mining techniques were discovered to analyze past medical history as a predictor and forecast potential medical costs [1]. It is, however, difficult to ensure a perfect model to achieve a perfect estimate. Researchers examined patient-level health care costs in both the short and long term and discovered a clear temporal link over various time spans [2]. The supervised learning methods for predicting healthcare costs using Artificial Neural Network (ANN) and the Ridge regression model using Empirical Evaluation produce approximations but struggle for time-series implementation [3]. Feature engineering was crucial in capturing temporal patterns in the data and reducing the number of features with minimal loss [4]. People have worked with comparative different approaches like supervised learning, Time-Series forecasting using statistical, neural, and ensemble architectures [5].

3 Dataset

This data was collected from the repository of the book where they simulated data using demographic statistics from the US Census Bureau which indicates approximately reflection of real-world conditions [6]. This dataset contains the age of the primary beneficiary (excluding those above 64 years, since they are generally covered by the government), policy holder's gender (it can be either male or female), the body mass index or BMI (weight (unit:kilograms) divided by height (unit:meters) squared), the number of children/dependents covered by the insurance plan, either smoker or non-smoker (it can be only two strings, either yes or no), and the beneficiary's place of residence in the US. A total of 1338 patients were enrolled in the dataset, with features reflecting patient characteristics as well as total medical costs paid to the plan for the calendar year. The structure of the CSV file that contains the dataset can be described as in Table 1.

Table 1: Description of the dataset features and feature values

Features	**Value type**	**Value description**
Age	An integer	: A number indicating the beneficiaries age less than 64 years
Sex	A string	: male or female
BMI	A float	: weight (in kilograms) divided by height (in meters) squared
Children	An integer	: the number of children/dependents covered by the insurance plan
Smoker	A string	: yes or no
Region	A string	: four geographic regions: northeast, southeast, southwest, or northwest

Aside from these features, the dataset contains another attribute that represents medical charges as the marker to train the dataset, indicating suitability for supervised model.

4 Data Processing

4.1 Categorical Data Transformation

In the dataset, we found the following categorical features: sex, smoker, and country. There was two categories for gender, two categories for smoker or non-smoker and four categories for the region. Since

these feature values are fixed and constrained, we transformed them and assigned them a numerical value beginning with zero.

4.2 Removing Outliers

To eliminate a moderate amount of noise from our dataset, we measured the interquartile range (IQR) and removed all data points that did not fall within the IQR. We regarded data points that exceeded the IQR as outliers, such as charges above 30000, which imply noise in the dataset. Table 2 shows the IQR difference for each feature.

Table 2: Mean of the Inter Quartile range $((Q_3 - Q_1)/2)$

Features	**IQR**	**Features**	**IQR**
Age	24.00	sex	1.00
BMI	8.40	children	2.00
smoker	0.00	region	1.00
charges	11899.62		

The IQR filter removed 283 records from a total of 1338, and the final dataset duration is 1055. The segment on exploratory data analysis represented the visual representation of outliers for individual features versus charges. The problem was supervised, the predicted output (y) could be compared with the given actual output (y). We also divided the dataset into two parts: training and testing sample. We separated the 15% dataset as the testing dataset and the rest of the dataset as training samples. After training the training set, we implemented developed models in the testing dataset.

4.3 Feature Combination and Replacement

The number of children in relation to age and BMI has an interconnected relationship, as shown in Figure 1 of exploratory data analysis. We looked at the correlation between them as well as the correlation with charges to analyze the connectivity power. We discovered that charges rise in direct proportion to age and the number of children, but fall in direct proportion to BMI. As a result of our results, we combined these three features with charges and addressed a new feature called stress level, which replaced the previously mentioned three features.

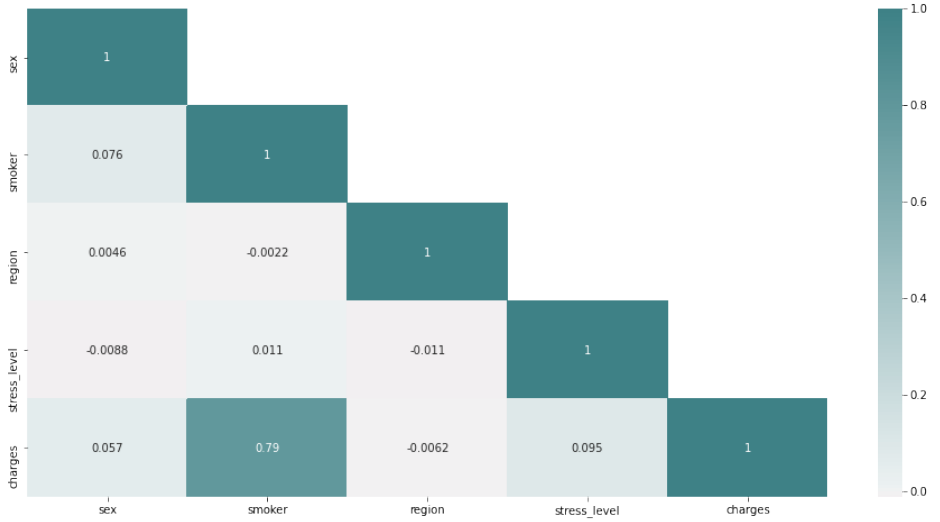


Figure 1: Correlation between individuals charges with other features including stress_level and excluding age, children, BMI

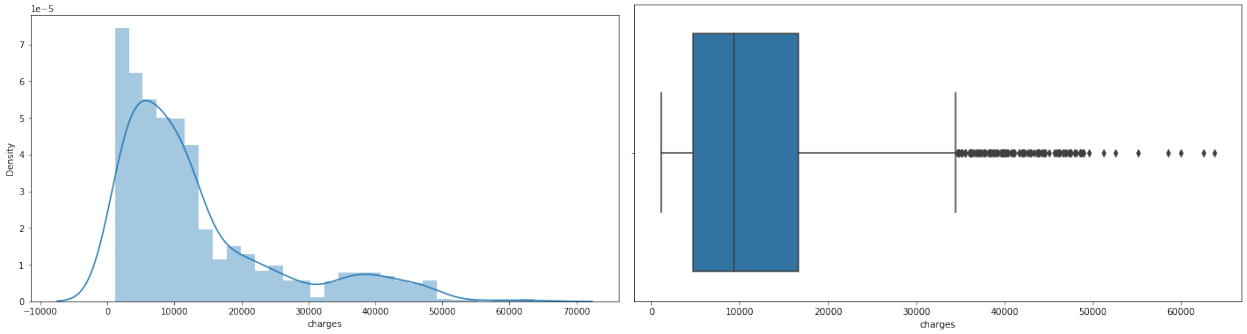
$$data[stress\_level] = \frac{data[children] * data[age]}{data[BMI]}$$

4.4 Normalization

On our dataset, we used min-max scaler normalization, which scales each individual feature in range (0,1). The aim of this normalization is to convert the values of the dataset's numeric columns to a standard scale without distorting variations in value ranges.

5 Exploratory Data Analysis

Firstly, to observe the distribution of the charges we plotted distribution for the values of *charges* feature and noticed that cost were most dense in the beginning and were gradually falling afterwards. Figure 2a indicates that most peoples insurance cost were between $15,000$ and very few peoples insurance cost were between $15,000$ to $50,000$. Also, the boxplot shows how high cost charges contribute to outliers in Figure 2b.

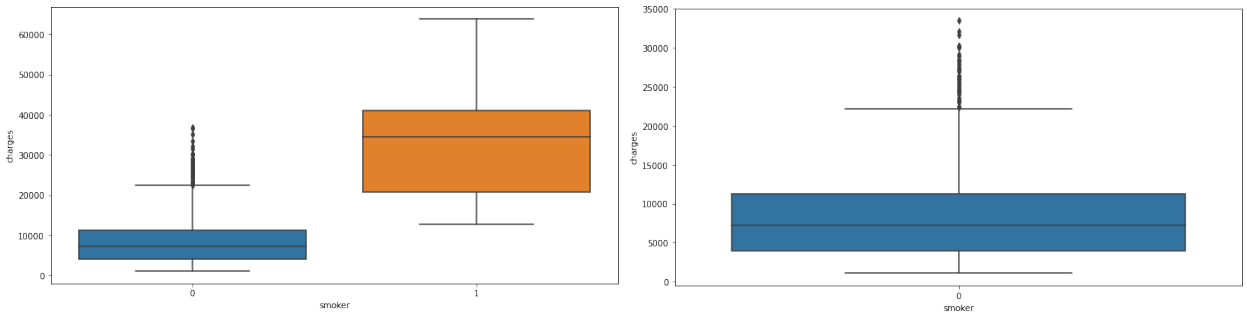


(a) Charges distribution per number of people

(b) Costly charge outliers

Figure 2: Charge distribution for different features

Figure 3a depicts the allocation of charges for smokers and non-smokers. We can see that smokers face a much wider range of charges than nonsmokers. The boxplot of smokers and nonsmokers with charges after eliminating outliers based on IQR is shown in 3b. However, a critical observation of this statistic is that if we consider the data points that were not in the IQR, then all data points for smokers are omitted. Since there were a small number of people who smoked but had a large medical bill, IQR causes an imbalance between the mean and an entire group is erased. This observation helped us to explore min-max scalar normalization.



(a) Charges distribution for smoker and non-smoker before removing IQR outlier

(b) Charges distribution for smoker and non-smoker after removing IQR outlie

Figure 3: charges distribution for smoker and non-smoker

Figure 4 shows correlations of charges with different features. It also shows that the highest correlation exists between smoker feature and charges. We can also see that the second and third strongly correlated features were age and BMI which led us to feature combination and replacement experiment referred in Figure 1.

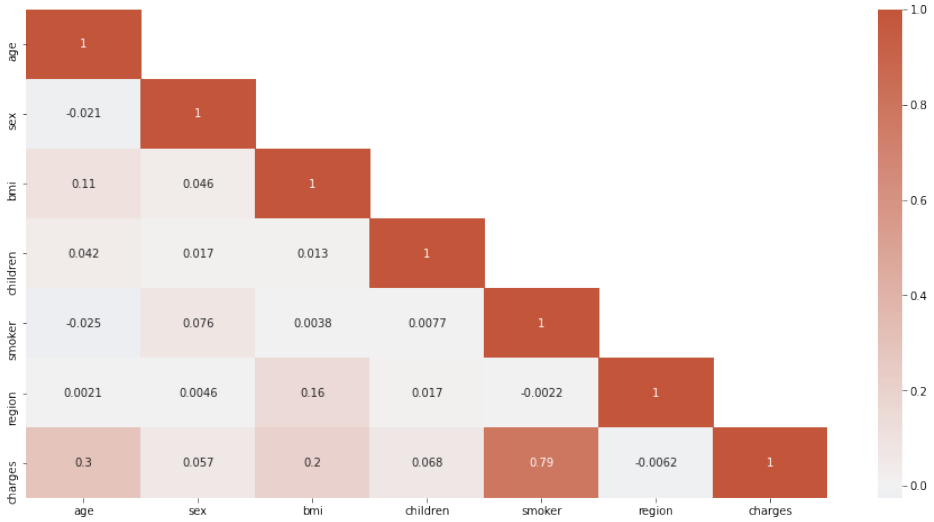


Figure 4: Correlation between individuals important features and medical costs

We also noticed from Figure 5 that another categorical feature namely region had a synchronized IQR in them and their mean were almost same.

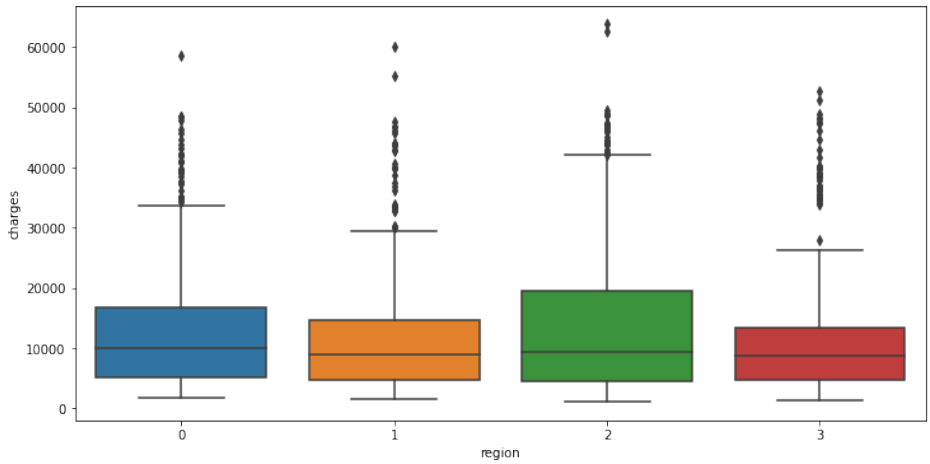


Figure 5: Distribution of charges in different regions after IQR outlier removal

6 Methodology and Experiment

6.1 Cost Functions

We used root mean square error (RMSE) as the cost function although we initially started with mean absolute error (MAE).

$$MAE = \frac{\sum_{i=1}^{n} |y_i - y_i^P|}{n}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - y_i^P)^2}{n}}$$

The RMSE is a quadratic scoring rule that calculates the average magnitude of the error. Since errors are squared before being averaged, the RMSE gives large errors a relatively high weight. As a result, the RMSE is most useful when large errors are particularly undesirable. When dealing with large error values, RMSE performs better in terms of representing efficiency. When lower residual values are preferred, RMSE is more useful. Since it does not square values, MAE is less than RMSE. As sample size increases, RMSE appears to be greater than MAE. MAE does not often penalize significant errors. When the conditional distribution of the observations is asymmetric and an unbiased fit is required, the conditional mean minimizes the RMSE and the conditional median minimizes the MAE. Low volume sales data, for example, usually have an asymmetric distribution. The optimization of the MAE, may make MAE-optimal forecast value a flat zero.

6.2 Supervised Learning Algorithms

From the three main categories of ML models: supervised, unsupervised, and semi-supervised models, our problem space denotes supervised model [7]. The dataset is called a training set in supervised learning and must contain all input variables and their corresponding output variables. We divided the dataset into two sections, training and testing, in a 17:3 ratio. The testing dataset has its output label which was used to compare between predicted and actual dataset. With this dataset, linear, ridge, SVM, random forest, DNN algorithms were explored. By studying these models, a better performing model with higher accuracy was developed implementing kfold cross validation.

6.2.1 Linear and Ridge Regression

To forecast insurance costs, we used standard linear regression on our dataset. As testing data, we randomly selected 15% of our dataset's data points. Linear Regression is a supervised machine learning algorithm that predicts continuous performance with a constant slope. It is used to estimate values within a continuous range rather than to categorize them. This model's output is assessed by measuring mean squared errors for both training and testing results. Finally, we compared our estimated charges to the actual charges of testing data points. Figure 6a shows how prediction and actual output differs from each other for linear regression and Figure 6b shows how how prediction and actual output differs from each other for ridge regression.

The train error and testing error was 6128.42 and 5558.94 accordingly. the testing error was less than the training error, indicating that no overfitting has occurred. However, in order to be more conclusive, we switched to ridge regression, also known as regularized linear regression. Cross validation is used to determine the value of the regularization parameter (α). We tried several different values of (α) in the range and used cross validation to find the one that works the best. Regularized linear regression produces less testing error than linear regression, confirming that there was no overfitting; however, we intended to explore a few more methods to get a better result.

6.2.2 Support Vector Regression

As SVM deals better in a smaller dataset and we failed to obtain a moderate performance for predicted results from linear and ridge regression models, We explored support-vector machines (SVM). SVM are supervised learning models with associated learning algorithms and for lesser training dataset, SVM handles outliers better. We followed linear kernal from linear and rbf kernal as linear SVMs (or logistic regression) for linear problems. The SVM algorithm is not appropriate for large data sets. When the data set contains more noise, i.e. target classes overlap, SVM does not perform well. The SVM can underperform when the number of features for each data point exceeds the number of training data samples. From Figure 6c, we can see that SVM did not perform well for our problem space. One reason could be that our testing dataset was only 17% of the training dataset. However, even after increasing the testing sample size, we obtained lower performance than with ridge regression.

6.2.3 Neural network

Neural network approaches are better for fault tolerance and predictive analysis as the hidden layers of neural network are used to improve prediction accuracy. We implemented our data set on a neural network and got the result with five layers where in four of them we used RELU activation function and on one of them we used linear activation. We used six input parameters. There were 128 neurons in the first hidden layer, 256 neurons in the second, third and fourth hidden layer, 1 neuron in the last layer. However, from the Figure 6d we can see that for the prediction from this model, we were getting the largest error.

6.2.4 Random forest

Random forests, also known as random decision forests, are an ensemble learning method for classification, regression, and other tasks that works by constructing a large number of decision trees during training and then predicting the class that is the mode of the classes or the average prediction of the individual trees. While increasing the trees, Random Forest adds more randomness to the model. When splitting a node, it looks for the best function among a random subset of features rather than the most appropriate feature. As a result, there is a wide range of variety, which leads to a stronger model in general. In our problem space it gave the best result for both training and testing sample although the error difference between training and testing sample was pretty higher. Figure 6e shows the testing sample performance by random forest with a moderate error. So from our realization of overfitting, we explored a few ways to manage overfitting such as polynomial feature transformation and kfold cross validation with hyperparameter tuning and its performance is shown on Figure 6f. All models performance on testing sample were reflected combinely on Figure 6. From the table 3, we can see that we got the best result so far from random forest regression.

Table 3: Performance of various models in terms of RMSE

Approach	**Training sample error**	**Testing sample error**
Linear	6128	5558
Ridge	6128	5559
SVM	12898	12795
Neural Network	15645	15746
Random Forest	1860	4674
Upgraded Random Forest	2739	3098

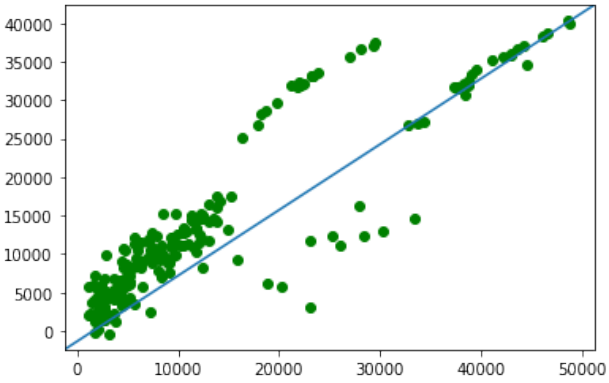
6.3 Polynomial Feature Transformation

When using random forest regression, we obtained a better scenario for the training dataset. However, it did not support the testing dataset as well as it did the training dataset, indicating an overfitting problem. And we dug deeper to find a solution to overfitting using polynomial feature transformation. However, we used a degree of 3 analysing bias-variance trade-ff.

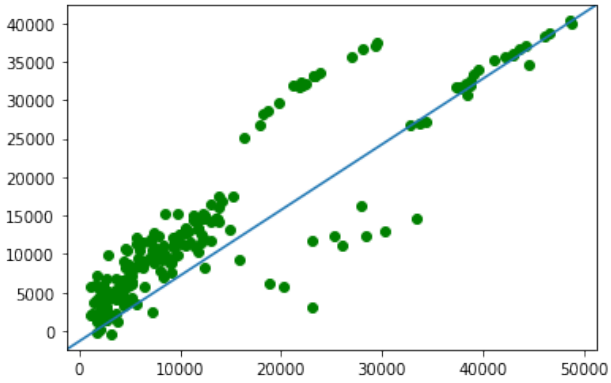
6.4 Hyperparameter tuning

The problem of selecting a set of suitable hyperparameters for a learning algorithm is known as hyperparameter optimization or tuning in machine learning. Grid search is arguably the most fundamental hyperparameter tuning technique. We simply create a model for each possible combination of all of the hyperparameter values given, evaluate each model, and choose the architecture that produces the best results. There is also another method known as the randomized search method. However, to fix the overfitting problem, we experimented with various hyper-parameter tuning for the random forest model using grid search technique.

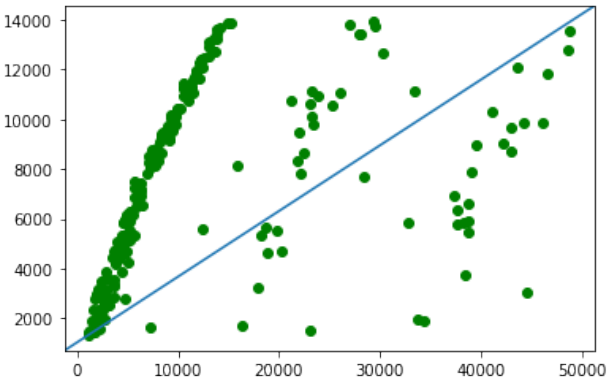
- **n_estimators:** The number of trees in the forest is estimated by n estimators. A higher value of it supports in better fitting. The default is 100.



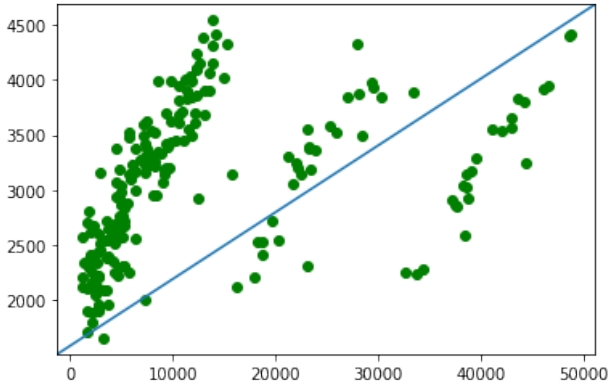
(a) Linear regression



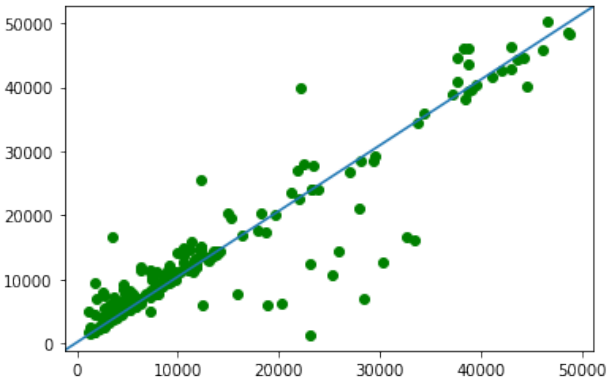
(b) Ridge regression



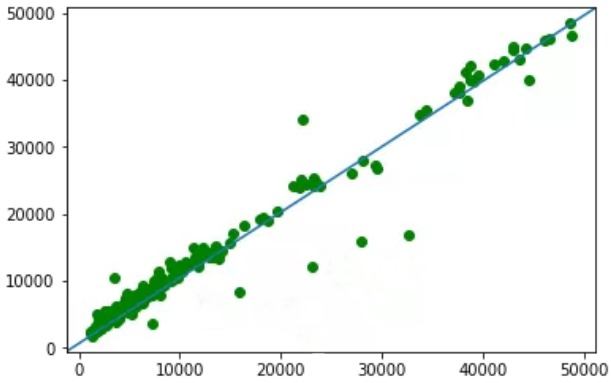
(c) Support vector regression



(d) Neural network



(e) Random Forest regression



(f) Improved random Forest regression

Figure 6: Actual vs. predicted y for different models

- **max_depth:** Each decision tree has a maximum number of levels. A higher value means better fit. The default value is "None," which means there is no depth limit.
- **min_sample_leaf:** A leaf can only have a certain number of data points. A lower positive value of it results in better fitting. The default value is two.
- **min_sample_split:** The minimum number of data points that can be put in a node before it is broken. The default value is two.
- **max_features:** The maximum number of features taken into account when breaking a node. Default value is "Auto" which means all features are considered.
- **bootstrap:** Method for sampling data points (with or without replacement).

We explored the effects of various hyperparameters on the prediction and in Figure 7 we observed that, we get the best performance for n_estimator's value of around fifty. For the other hyperparameters, almost all value has similar performance. How and for what n_estimator value from 1 to 200 effects on performance on testing and training sample is illustrated in Figure 7 and n_estimator value change was shown in x axis.

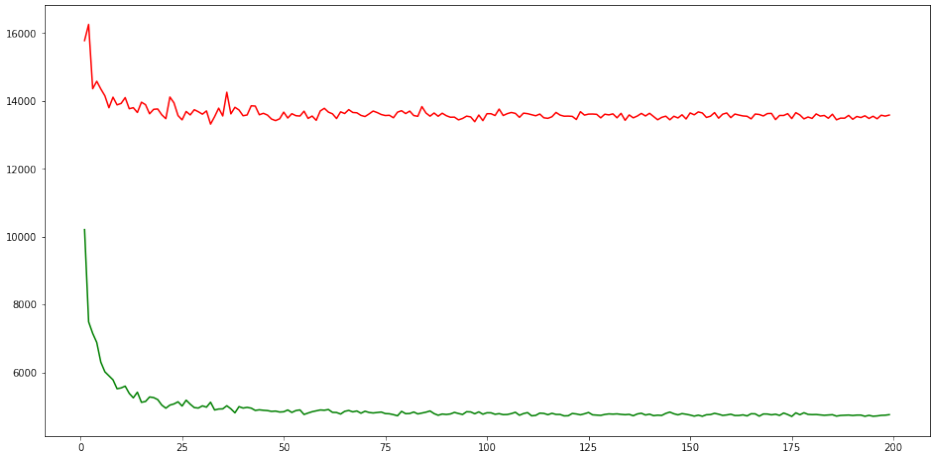


Figure 7: Different n_estimator's value reflected on RMSE error

Later, We tried different values of these parameters one at a time and get the best hyperparameters. We prefered grid search as it searches all possible combinations of the hyperparameters and find the best one. However, randomized search uses only the ones that we instruct it to. So, grid search is more computation cost intensive than randomized search. That being said, it serves a greater purpose.

6.5 Bias-Variance Analysis

In our project, when we first applied random forest regression on our training and testing dataset, we got training sample error of 1860 and testing sample error of 4674. So our model was overfitting for training dataset and we assumed it had high variance. From this understanding, we applied feature transformation with degree of 5 to fit our model in testing sample. However, we got training sample error of 4256 and testing sample error of 5109 which indicated it had high bias because it was not even fitting the training set then. We finally implemented polynomial feature transformation with a degree of 3 and cross validated the hyperparameters and it brought a bia-variance balanace in our model.

6.6 k-fold Cross-Validation

Cross-validation is a re-sampling technique used to test machine learning models on a small sample of data. That is, to use a small sample to approximate how the model would do in general when used to

make predictions on data that was not used during the model's training. The original sample is randomly partitioned into k equal sized sub-samples in k-fold cross-validation. In our model, hyperparameter results are then validated using 5-fold cross validation.

7 Conclusion

To predict insurance cost of an individual, smoking habits affect the cost estimation the most. Our available information size affected out final result. If we could get more data we could predict medical cost more accurately with our improved model.

References

[1] D. U. Himmelstein, D. Thorne, E. Warren, and S. Woolhandler, "Medical Bankruptcy in the United States, 2007: Results of a National Study," *The American Journal of Medicine*, vol. 122, pp. 741–746, Aug. 2009.

[2] C. Yang, C. Delcher, E. Shenkman, and S. Ranka, "Machine learning approaches for predicting high cost high need patient expenditures in health care," *BioMedical Engineering OnLine*, vol. 17, p. 131, Nov. 2018.

[3] M. A. Morid, K. Kawamoto, T. Ault, J. Dorius, and S. Abdelrahman, "Supervised Learning Methods for Predicting Healthcare Costs: Systematic Literature Review and Empirical Evaluation," *AMIA Annual Symposium Proceedings*, vol. 2017, pp. 1312–1321, Apr. 2018.

[4] A. M. Jödicke, U. Zellweger, I. T. Tomka, T. Neuer, I. Curkovic, M. Roos, G. A. Kullak-Ublick, H. Sargsyan, and M. Egbring, "Prediction of health care expenditure increase: how does pharmacotherapy contribute?," *BMC Health Services Research*, vol. 19, p. 953, Dec. 2019.

[5] S. Kaushik, A. Choudhury, P. K. Sheron, N. Dasgupta, S. Natarajan, L. A. Pickett, and V. Dutt, "AI in Healthcare: Time-Series Forecasting Using Statistical, Neural, and Ensemble Architectures," *Frontiers in Big Data*, vol. 3, 2020.

[6] B. Lantz, *Machine Learning with R*. Birmingham: Packt Publishing, Oct. 2013.

[7] S. Chibani and F.-X. Coudert, "Machine learning approaches for the prediction of materials properties," *APL Materials*, vol. 8, p. 080701, Aug. 2020.