Abstract

We have chosen a dataset extracted from Kaggle containing various characteristics of 1338 beneficiaries of a particular health insurance based in the United States (Choi 2018). Our study aims to determine which prediction models are most effective at predicting the medical charges claimed by the beneficiaries of the health insurance. In doing so, we have executed the machine learning pipeline consisting of problem definition, data collection, data preprocessing, data segregation, model training and model evaluation. Upon data preprocessing and segregation, we built four prediction models, namely, Linear Regression (LR), Polynomial Regression (PR), Support Vector Machine for Regression (SVR) and Random Forest Regressor (RFR). In order to achieve optimal model performance, we also performed hyperparameter tuning. Lastly, we evaluated the performance for each model by comparing them against a Naïve baseline and found that the best performing model for our dataset was RFR.

Introduction

As per our chosen dataset, we wish to perform prediction on a continuous target variable, namely, health insurance claims on medical charges based on multiple independent variables (e.g., age, sex, body mass index (BMI), number of children, smoker status and region). Accordingly, this is a problem that can be solved through regression analysis (Mitchell 2019). Consequently, we have built four different models in order to perform our prediction. LR examines the linear relationship between the independent and target variables in order to discover the best fitting line that can be used to make predictions. It works well where the data has a linear relationship (Mitchell 2019). On the other hand, PR, SVR, and RFR are better equipped at dealing with data that have non-linear relationships. PR uses LR with polynomial features to approximate non-linear functions in order to fit non-linear data points onto the line more accurately (Pant 2019). SVR operates by finding a hyperplane (in multidimensional space) between different data classes whereby future values are predicted based on the maximum number of points that are fitted on the hyperplane which are also within the decision boundary line. This ensures that we have the least error rate and provides a better fitting model (Sethi 2020). RFR constructs a multitude of decision trees and compute the mean prediction of the individual trees when generating the predicted output (Gurucharan 2020). Through testing each of these models, we hope to find the most effective model for predicting the medical charges from our dataset.

Materials and Methods

Continuing with the data preprocessing step from the machine learning pipeline, we begin with analysing the dataset by studying its variables and addressing any anomalies prior to feeding them to the model. These include but are not exhausted to checking for duplicated/missing values, outliers, invalid observations and/or any other irregularities (e.g., spelling/format errors). Based on our analyses of the dataset, we did not find any missing values, outliers nor invalid observations and/or any other irregularities. However, there was one duplicated value which was subsequently removed. As our dataset contains three categorical variables (eg., sex, smoker status and region), we assigned dummy numerical values in order to include and feed these variables into our models. This resulted in an increase in the dimensionality of our dataset from six to nine independent variables. To address this, we performed statistical analysis in order to understand the relationships between the target and independent variables and determine whether they need to be included/excluded from the models. In doing so, we used Pearson's correlation in order to draw a heatmap where we found that only the 'smoker' variable was highly correlated with the target variable whereas the others are not. This is further corroborated when we calculated the 'feature importance score' in order to determine the most to least important variables for the RFR model when making a prediction (Brownlee 2020). Nevertheless, due to the fact that we only have 9 variables, we decided to include all of them in building our models in order to avoid underfitting our models. The dataset is then split into training and testing sets. We will use the training set in order to train the model and perform hyperparameter tuning and use the testing set to evaluate our models. Before fitting the training set for each of the models to learn, we ensure that the data is normalised in order to change the values of the variables in the dataset to a common scale without losing information. This is to avoid the variables with higher values from intrinsically influencing the results due to its larger value (Jaitley 2018). Unlike the other models, we also performed Principal Component Analysis (PCA) on our LR model in order to reduce the dimensionality of the dataset from nine to seven components since LR is prone to overfitting (GeeksforGeeks 2020a). We then sequentially apply normalization (for all our models except RFR) and PCA (only for our LR model) followed by our respective predictive models and store them in a pipeline in order to automate the machine learning process (Sareen 2018). Upon fitting our respective models to the training set, we proceeded to tune the hyperparameters in order to choose a set of optimal hyperparameters that can provide the best prediction. As seen in Table 1 below, we do so through an iterative process whereby we define a range of possible values for all the hyperparameters, evaluate the performance of each model and select the hyperparameters that eventually produces the best results (Jordan 2017).

Model type	Hyperparameter	Range of hyperparameters	Optimal hyperparameters
LR	Not applicable		
PR	degree	1 - 10	2 (Default)
	interaction_only	True; False	False (Default)
SVR	kernal	Linear; RBF	RBF
	C	1000 – 100,000	50,000
	epsilon	0 - 10	5
	gamma	0 - 10	0.1
RFR	n_estimators	80 – 1000	105
	max_depth	10 – 110	95
	min_samples_split	1 – 20	14
	min_samples_leaf	1 - 20	12

Table 1. Hyperparameter tuning of predictive models

We evaluate the performance for each model using the test set by first establishing a Naïve baseline which provides a meaningful reference to compare the accuracy of our prediction models based on their root mean squared error (RMSE). RMSE is a good measure of how accurately the model predicts the target variable as it indicates how close the real values are to the model's predicted values. The goal is for the RMSE value of our models to be lower than the baseline thus indicating a better fit (Martin 2013). In selecting our best model, we then analysed each of their performance using the test set by comparing their RMSE against our baseline.

Results and Discussion

Model type	**Model pipeline**	**RMSE (without tuning)**	**RMSE (with tuning)**	**% improvement over baseline**
Naïve baseline	N/A	13172.95940839611*		N/A
LR	(StandardScaler(), PCA(n_components=7, random_state=1, whiten=True), LinearRegression())])	6399.7020215422845*		51.42%
PR	(StandardScaler(), PolynomialFeatures(),LinearRegression())])	4694.28273543746**		64.36%
SVR	StandardScaler(), SVR(C=50000, epsilon=5, gamma=0.1))])	13918.387312953908	4659.447714493853	64.63 %
RFR	RandomForestRegressor(max_depth=95, min_samples_leaf=12, min_samples_split=14, n_estimators=105, random_state=0))]	4764.678877424955	4375.409489171308	66.78 %

Table 2. Comparison of model performance on test set against baseline based on RMSE

**We did not perform hyperparameter tuning on Naïve baseline and LR as they do not have hyperparameters.*

*** Hyperparameter tuning was performed on PR but it was found that the default values were also the optimal values.*

As seen in Table 2, our best performing model is RFR, followed by SVR, PR and then LR. All of our models performed better after hyperparameter tuning apart from LR which does not require any hyperparameter tuning. It is not surprising that LR is our worst performing model as it oversimplifies the problem by assuming a linear relationship between the variables (GeeksforGeeks 2020a). Since most of our independent variables have a non-linear relationship with the target variable, PR, SVR and RFR are better equipped at handling these relationships (GeeksforGeeks 2020b). This would explain for their superior performance and the fact that their RMSE scores are close to one another. However, in looking at Figure 1 below, we can see that both PR and SVR can predict very accurately where the value of the charges is lower but becomes less accurate as the value increases. This is likely due to the fact that there are other factors such as an individual with pre-existing condition/illness that can result in higher medical charges. Since we do not have this information, it is difficult for the models to predict. On the other hand, RFR is able to predict with good accuracy on both the lower and higher values. This is because RFR performs better with categorical variables which is the majority of our variables (Au 2018). It does so by dealing with the lower and higher values separately by forming a multitude of decision trees and calculating the average as the final value for the target. From a commercial standpoint, an insurance company would be more interested in predicting the higher charges as it would greatly influence their profits. Since insurance companies are not privy to personal medical records of their customers (Australian Government n.d.), it would be more beneficial to employ RFR as our model. While SVR and PR are well fitted for predicting lower charges, their performance is limited by the fact that they do not handle categorical variables or predict the higher charges as well as RFR.

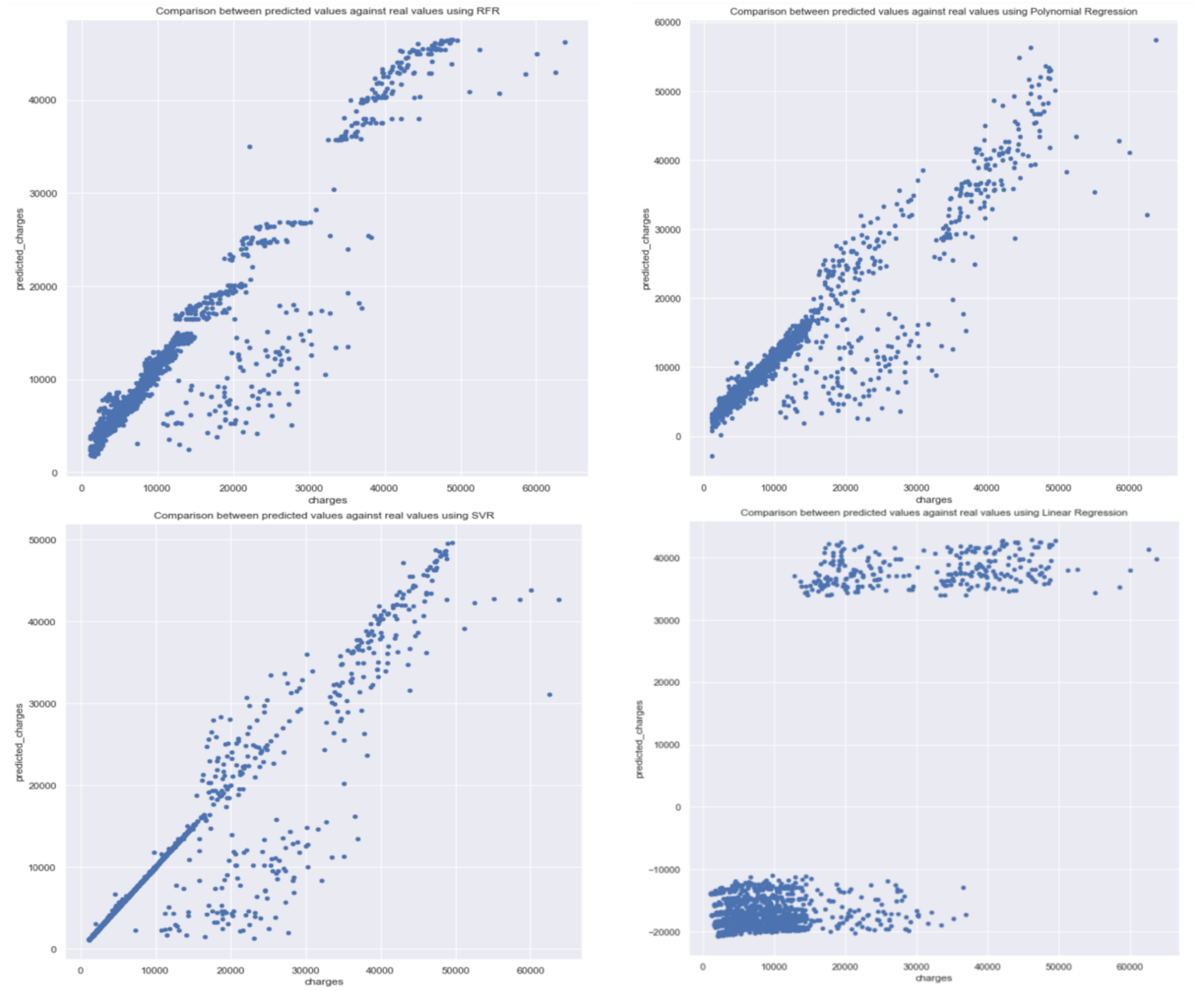

Figure 1. Comparison between predicted charges against actual charges using RFR (top left), PR (top right), SVR (bottom left) and LR (bottom right).

Conclusions

RFR is a useful technique as it helps us to deal with the limitations pertaining to making predictions using categorical variables. It also performs well on variables with linear or non-linear relationships. For our particular dataset, we observed that RFR is superior to the other models as a prediction method.

LR cannot capture the nonlinearity in the dataset. Although PR and SVR is adept at dealing with non-linear relationships, they are unable to predict the higher charges as accurately as RFR since they are not as good at handling categorical variables. While the performance of our models could be greatly improved if we had the health records for each of the beneficiaries, it is information that cannot be readily obtained due to privacy concerns. Nevertheless, there may be other factors that can influence medical charges that we have not considered. As future work, we should seek advice from domain experts within the insurance industry in order to determine the types of data that is not only relevant to improving our models, but are also accessible without breaching any privacy laws or concerns.

References

Australian Government n.d., *My health records: frequently asked questions*, Australian Government, viewed 16 February 2021, <https://www.myhealthrecord.gov.au/for-you-your-family/howtos/frequently-asked-questions#:~:text=Only%20healthcare%20provider%20organisations%20involved,access%20your%20My%20Health%20Record.>

Au, TC 2018, 'Random forests, decision trees, and categorical predictors: the "absent levels" problem', *Journal of Machine Learning Research*, vol. 19, pp. 1-30.

Brownlee, J 2020, *How to calculate feature importance with python,* Machine Learning Mastery, viewed 15 February 2021, < https://machinelearningmastery.com/calculate-feature-importance-with-python/>

Choi, M 2018, *Medical cost personal datasets*, Kaggle, viewed 15 February 2021, < https://www.kaggle.com/mirichoi0218/insurance>

GeeksforGeeks 2020a, *Advantages and disadvantages of linear regression*, GeeksforGeeks, viewed 16 February 2021, <https://www.geeksforgeeks.org/ml-advantages-and-disadvantages-of-linear-regression/>

GeeksforGeeks 2020b, *Advantages and disadvantages of different regression models,* GeeksforGeeks, viewed 16 February 2021, <https://www.geeksforgeeks.org/advantages-and-disadvantages-of-different-regression-models/?ref=rp>

Gurucharan, MK 2020, *Machine learning basics: random forest regression,* Towards Data Science, viewed 15 Februrary 2021, < https://towardsdatascience.com/machine-learning-basics-random-forest-regression-be3e1e3bb91a>

Jaitley, U 2018, *Why data normalization is necessary for machine learning models*, Medium, viewed 16 February 2021, < https://medium.com/@urvashilluniya/why-data-normalization-is-necessary-for-machine-learning-models-681b65a05029>

Jordan, J 2017, *Hyperparameter tuning for machine learning models,* Jeremy Jordan, viewed 16 February 2021, < https://www.jeremyjordan.me/hyperparameter-tuning/>

Martin, KG 2013, *Assessing the fit of regression models,* The Analysis Factor, viewed 16 February 2021, < https://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/>

McKenzie, M 2019, *Selecting the correct predictive modeling technique,* Towards Data Science, viewed 15 February 2021, <https://towardsdatascience.com/selecting-the-correct-predictive-modeling-technique-ba459c370d59#:~:text=Linear%20regression%20is%20to%20be,and%20dependent%20variables%20are%20linear.>

Pant, A 2019, *Introduction to linear regression and polynomial regression*, Towards Data Science, viewed 15 February 2021, <https://towardsdatascience.com/introduction-to-linear-regression-and-polynomial-regression-f8adc96f31cb>

Sareen, S 2018, *Pipelining in python,* Medium, viewed 16 February 2021, < https://medium.com/@shivangisareen/pipelining-in-python-7edd2382f67d>

Sethi, A 2020, *Support vector regression tutorial for machine learning,* Analytics Vidhya, viewed 15 February 2021, < https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/>

Prepared by:
Nazzeef Nazri
a1621410