

**Abstract:**

For insurance companies, a very important task is to determine an ideal premium for every insured individual given a few independent variables such as age and body mass index. Using historical data, we ’ll try to construct a statistical model to compute the expected charges for each individual.

**INDEX**

1. **EXECUTIVE SUMMARY...…………………………………………………………………2**
   1. Brief Introduction
   2. Solution & Approach
   3. Results & Conclusions
2. **EXPLORATORY DATA ANALYSIS………………………………………………………4**
   1. Distribution of target variable
   2. Correlation between the variables
   3. Age vs Charges
   4. BMI vs Charges
   5. Distribution of Charges by Region
   6. Charges distribution by Age, Sex and Region
   7. Outliers
3. **MODEL DEVELOPMENT.…………………………………………………………….….10**
   1. Feature Engineering
   2. Linear Regression Model
   3. Random Forest
   4. Support Vector Regression
   5. Gradient Boosting
   6. XG Boost
   7. Model Comparison
4. **CONCLUSION……………….…………………………………………………………….….20**
5. **Executive Summary**
   1. Brief Introduction:

Medical insurance companies, are often looking for analysis to determine an ideal premium for every insured individual given a few independent variables like Body Mass Index, dependents, smoking habits etc. Early identification of those with a higher risk of hospitalization could help in making efficient analysis to determine the charges.

Using historical data, the project will try to construct a statistical model to compute the expected charges for everyone. Before that, exploratory graphs and quick modelling will give us basic insights.

* 1. Solution & Approach:

There are no issues with the data, such as missing values, so there is no need for data pre-processing or feature engineering. We performed the exploratory data analysis and analysed the correlation between the target variable and other predictor variables.

We split the dataset as 80% training data and 20% testing data. We developed the Linear Regression Model, Random Forest Model, XG Boost Model, Gradient Boosting Model and Support Vector Regression Model to predict the insurance charges. Then compared the in-sample and out of sample prediction accuracy for all the models.

* 1. Results & Conclusions:

To conclude out of the many supervised learning models using this project, XGBoost has the highest out-sample performance with highest R-square. Other models like Gradient Boosting was also quite close with almost similar out-of-sample R-squared value. This can be stated to the fact, that XG Boosting uses a more regularized model formalization to control over-fitting, which gives it better performance. Additionally, it has extra randomization parameter that can be used to reduce the correlation between trees.

Insurance charges can be best predicted using XG Boosting using the predictors in the model formula. For a business perspective, insurance charges can be predicted using information such as age, bmi30: smokeryes, etc for individual customers. This could also derive customer loyalty as the insurance charges suits their income and category.

* 1. References:

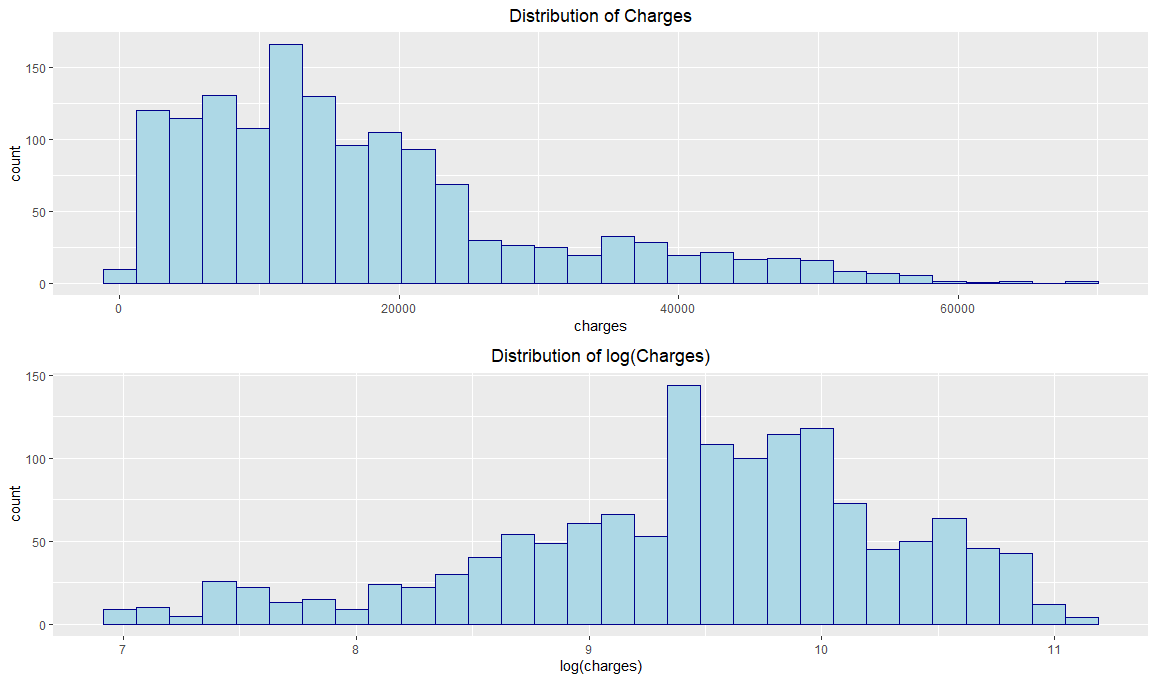
1. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani (2014). An Introduction to Statistical Learning with Applications in R. Springer New York, ISBN-13: 978-1-4614-7138-7.
2. **Dataset:** <https://www.kaggle.com>
3. **Exploratory Data Analysis**

* Dataset has 1429 observations of below 7 variables
* Charges is our target variable
* All the variables are normally distributed, with notable kurtosis in age, sex, region and charges columns

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Age​** | **sex​** | **bmi​** | **children​** | **smoker​** | **region​** | **charges​** |
| Int​ | Factor​ | Num ​ | Int​ | Factor​ | Factor​ | Num​ |

Variables in the dataset

* **age**: Age of the primary beneficiary
* **sex**: Insurance contractor gender; female, male
* **bmi**: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
* **children**: Number of children covered by health insurance / Number of dependents
* **smoker**: Whether smoker or not
* **region**: The beneficiary's residential area in the US, northeast, southeast, southwest, northwest
* **charges**: Individual medical costs billed by health insurance
  1. Distribution of target variable
* From the below histogram, we can observe that distribution of Charges is skewed
* With Log transformation the distribution is much more symmetric



Distribution of Charges

* 1. Correlation between the variables
* Correlation between the variables are very low
* A strong correlation can be observed between Charges and Smokers
* Age, BMI and Sex are evenly distributed
* Many respondents have 0 children or 0 dependents
* Some other notable significance, even if less with target variable are - Age with charges and bmi with age

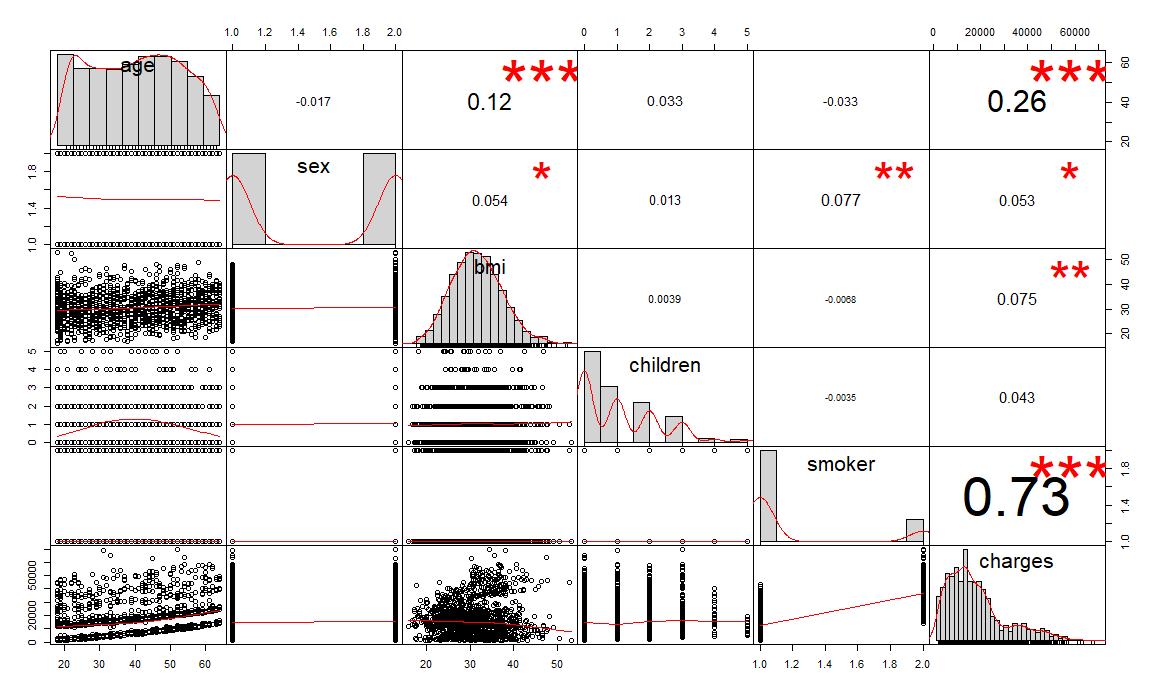


Fig: Correlation Plot

* 1. Age vs Charges
* Age is positively corelated with charges
* Insurance charges paid by smokers has high variance, compared to non-smokers of same age
* It can be observed from the below Age vs Charges correlation plot that irrespective of age, non-smokers pay less insurance charges
* To incorporate this observation, in this analysis age is separate by smoking status i.e., Smokers vs Non-Smokers

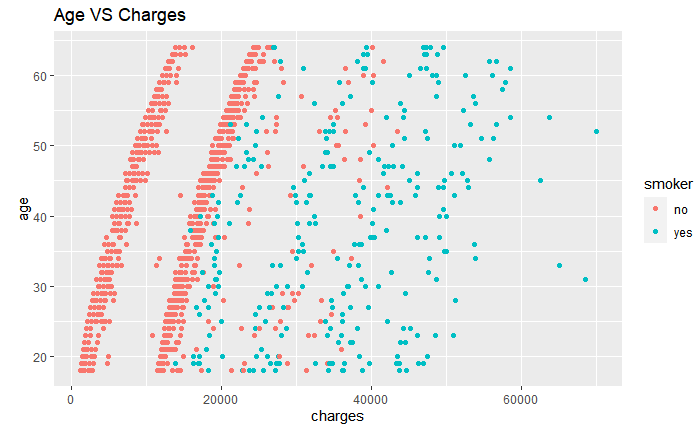


Fig: Correlation between Charges and Age

* 1. BMI vs Charges
* Charges for smokers with bmi over 30 are high
* For non-smokers, the charges are not correlated with BMI

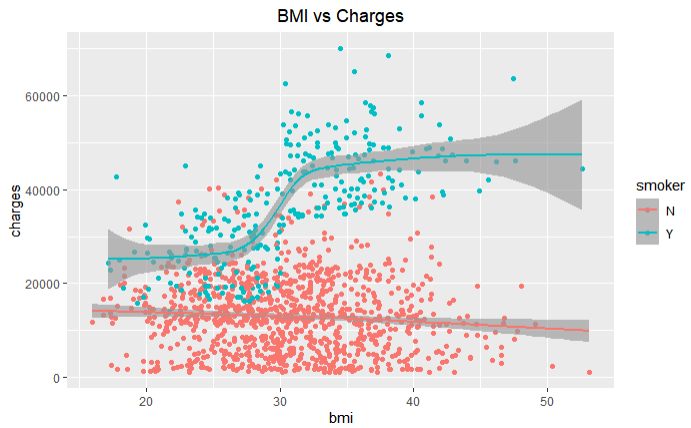


Fig: Correlation between Charges and BMI

* 1. Distribution of Charges by Region
* Due to some specific regions, overall charges are skewed to the right
* NE and NW regions have high mean which is pulling the overall distribution towards the right, indicating higher income group or expensive premium

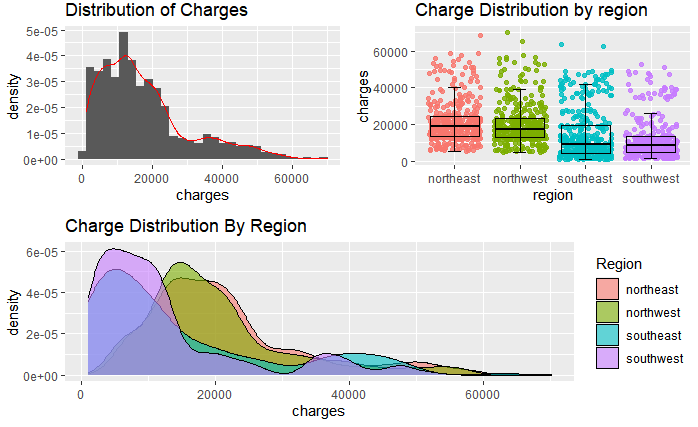


Fig: Distribution of Target Variable

* 1. Charges distribution by Age, Sex and Region
* Age is broken down into 3 groups:
* Youth 18 to 35 years
* Mid Aged 36 to 50 years
* Old 51 to 80 years
* For each region and age group the distribution of charges is analysed separately for Male and Female populations.
* It is observed that NE and NW regions have higher charges for males in the old category whereas Females have higher charges in the Mid Age
* Female youths in the SE and SW have lesser mean charges

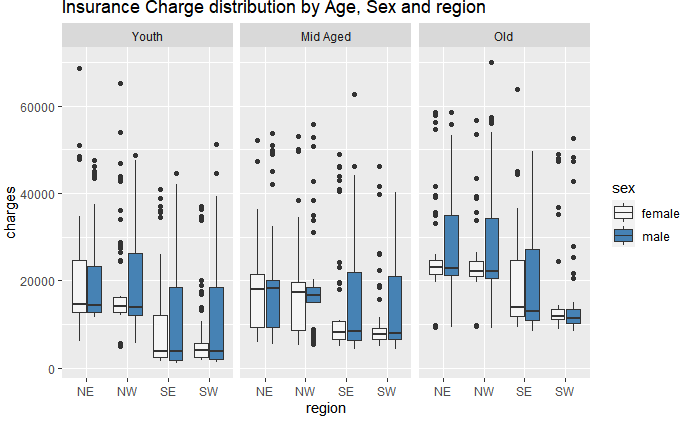


Fig: Charges distribution by Age, Sex and Region

* 1. Outliers
* Following groups were having higher outliers in their charges,
* Insurance payers at age 30 and 40
* Insurance payers with 1 and 3 children / dependents
* Insurance payers from NE and NW regions

1. **Model Development**
   1. Feature Engineering

* Region, Sex and Smokers were converted as factors before building the model
* BMI30 variable introduced as a factor with level as “Obese” for BMI >30, else “Not Obese”
* The dataset is split as 80% training sample and 20% testing sample
* The target variable in both the train and test data were normalized
  1. Linear Regression Model

Developed a simple Linear Regression Model. First the linear regression model was developed without any interaction effects and obtained as Adjusted R Square value of 0.701. In the Exploratory Data Analysis as we have already observed that there is a high correlation between the BMI and Smoker predictor variables. So, a second linear regression model was developed with introduction of the interaction effect between BMI (>30) and Smoker (Yes). Now the Adjusted R Square value improved to 0.736, thus the introduction of the interaction term improved the model more than would be expected by chance.

|  |  |  |  |
| --- | --- | --- | --- |
| INTERACTION EFFECT​​ | FORMULA​​ | ADJ. RSQUARED​​ | SIGNIFICANT PREDICTORS​​ |
| NO​​ | charges ~ age + smoker + bmi30 + sex + children + region​​ | 0.701 | Age, children, smokeryes,bmi30obese​​ |
| YES​​ | charges ~ age + smoker + bmi30 + sex + children+ region +**bmi30\*smoker​**​ | 0.736 | Age, children, smokeryes,bmi30obese, region SW |

Table: Linear Regression Model Summary

To measure the accuracy of the final linear model, the target variable from the in-sample data was first predicted. To improve the prediction efficiency, statistical models like Random Forest, XG Boost, Gradient Boosting, SVM: Regressor will be explored and compared. The model formula with BMI and Smoker interaction will be used for all the models.

* 1. Random Forest

As Random Forest models can give significantly higher accuracy in prediction, next a random forest model was developed on the train dataset using the model formula with interaction.

To evaluate the accuracy of the random forest model and to select appropriate values for tuning parameters, such as the number of candidate predictors that are randomly drawn for a split, OOB Error (Out of Bag Error) plot with different mtry is plotted.

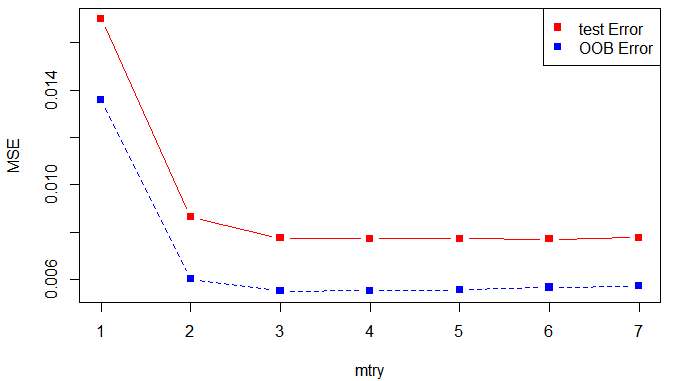


Fig: Random Forest OOB Error Plot

From the OOB Error plot, it can be observed that with mtry 2, the test error and OOB error are lesser. The random forest model was built with 2 mtry and 200 ntree.

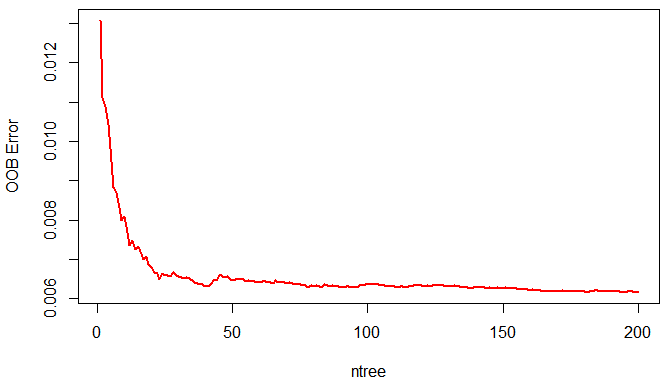


Fig: OOB Error vs Number of trees

From the below variable importance plot from the developed random forest, it can be observed that smoker status has the highest importance and then region, age and BMI follow it in predicting the insurance charges

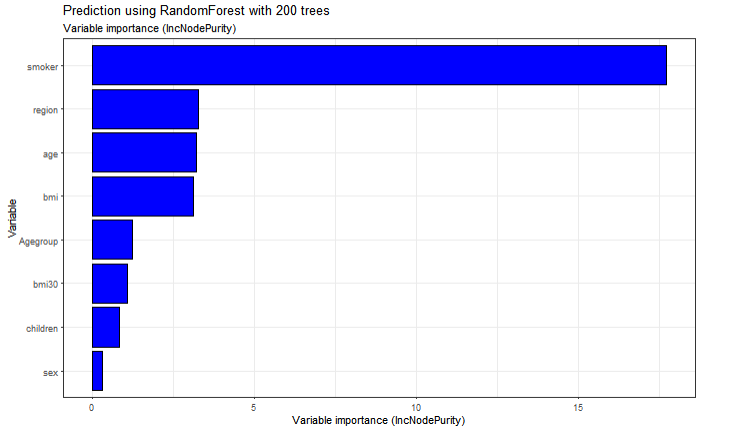


Fig: Variable Importance Plot

To measure the accuracy of the random forest model, the target variable was predicted for the in-sample data. The predicted ‘charges’ had a RMSE of 0.05219575. Then for the out of sample data the predicted ‘charges’ had a RMSE of 0.10141311. Thus, the out of sample RMSE has slightly decreased compared to Linear Regression.

|  |  |  |  |
| --- | --- | --- | --- |
|  | RMSE | R.squared | MAE |
| Training Set | 0.05219575 | 0.9317544 | 0.03353768 |
| Testing Set | 0.10141311 | 0.7818169 | 0.06360183 |

Fig: Random Forest Model Summary

* 1. SVM: Regressor

Support Vector Regression gives us the flexibility to define how much error is acceptable in our model and will find an appropriate line (or hyperplane in higher dimensions) to fit the data.

In contrast to Ordinary Least Square, the objective function of SVR is to minimize the coefficients — more specifically, the *l*2-norm of the coefficient vector — not the squared error. The error term is instead handled in the constraints, where we set the absolute error less than or equal to a specified margin, called the maximum error, ϵ (epsilon). We can tune epsilon to gain the desired accuracy of our model.



Fig: SVM Best Tune Parameters

To measure the accuracy of the SVM model, the target variable was predicted for the in-sample data. The predicted ‘charges’ had a RMSE of 0.06807194. Then for the out of sample data the predicted ‘charges’ had a RMSE of 0.09972658. Thus, the out of sample RMSE has slightly decreased compared to Random Forest.

|  |  |  |  |
| --- | --- | --- | --- |
|  | RMSE | R.squared | MAE |
| Training Set | 0.06807194 | 0.8688119 | 0.03336189 |
| Testing Set | 0.09972658 | 0.7827709 | 0.05623169 |

Fig: SVM Model Prediction Summary

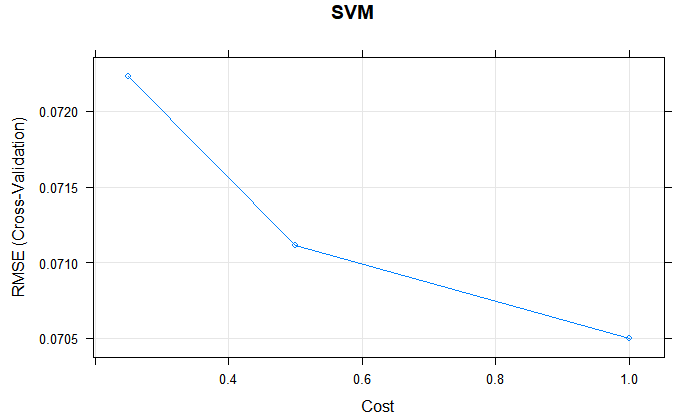


Fig: SVM Cost Plot

* 1. Gradient Boosting

Gradient boosting is an approach where new models are created, which predicts the residuals or errors of the prior models and then combined to give the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models. This approach supports both regression and classification predictive modelling problems.

Gradient Boosting comprises of the following parameters,

* max\_depth - the max. depth to which intermediate trees can grow
* maximum number of iterations
* learning rate - specifies at what factor the tree values will be scaled when it builds upon the errors of previous intermediate trees

|  |  |  |  |
| --- | --- | --- | --- |
|  | RMSE | R.squared | MAE |
| Training Set | 0.06632627 | 0.8747767 | 0.03781746 |
| Testing Set | 0.09902834 | 0.7786041 | 0.05805065 |

Fig: Gradient Boosting Prediction Summary

To measure the accuracy of the XG Boost model, the target variable was predicted for the in-sample data. The predicted ‘charges’ had a RMSE of 0.06632627. Then for the out of sample data the predicted ‘charges’ had a RMSE of 0.09902834. Thus, the out of sample RMSE has slightly decreased compared to SVM Regression.

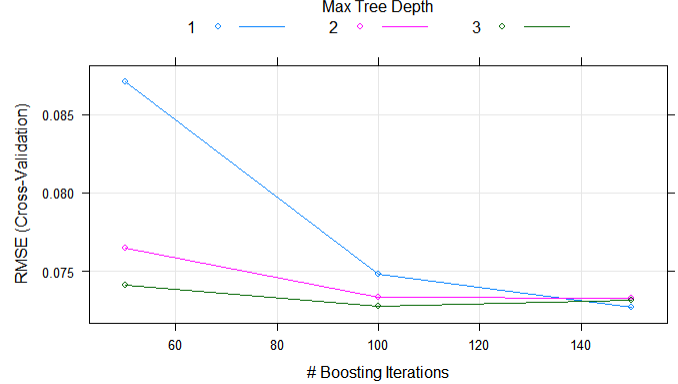


Fig: Gradient Boosting with Cross Validations

From the above cross validation plot, we can observe that with interaction depth 3 and 100 iterations, the RMSE (Root Mean Square Error) can be significantly lowered.

* 1. XG Boost

To further improve the prediction accuracy, XG Boost is used. Boosting is an ensemble technique where new models are added to correct the errors made by existing models. Models are added sequentially until no further improvements can be made.

The XG Boost model was developed trained on the train dataset with “Cross Validation” as resampling method and ‘RMSE’ as the metric to select the optimal model.



Fig: XG Boost Best Tune

From the XG Boost model feature importance plot it can be observed that Smokers, BMI30 – Obese, age, region SE and SW has the higher importance in the predicting the insurance charges.

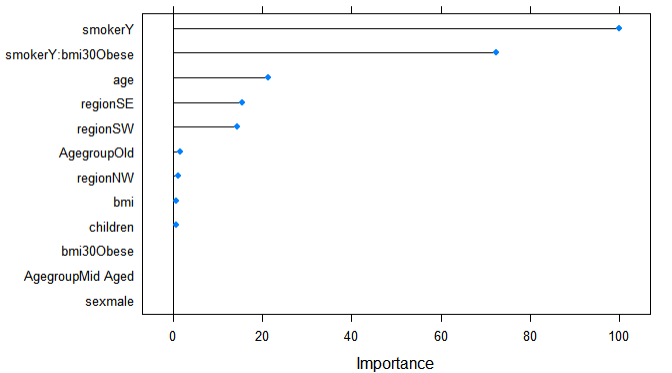


Fig: XG Boost – Model Feature Importance

To measure the accuracy of the XG Boost model, the target variable was predicted for the in-sample data. The predicted ‘charges’ had a RMSE of 0.06627407. Then for the out of sample data the predicted ‘charges’ had a RMSE of 0.09847384. Thus, the out of sample prediction accuracy has increased compared to gradient boosting.

|  |  |  |  |
| --- | --- | --- | --- |
|  | RMSE | R.squared | MAE |
| Training Set | 0.06627407 | 0.8749223 | 0.03744300 |
| Testing Set | 0.09847384 | 0.7852130 | 0.05784156 |

Fig: XG Boost Model – Prediction Summary

* 1. Model Comparison

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Seed | Model | Split |  | RMSE | R.SQUARED | MAE |
| 3238 | Random Forest | 70-30 | Training Set | 0.05217343 | 0.9316386 | 0.03307625 |
| Testing Set | 0.08453311 | 0.8135702 | 0.05331898 |
| 80-20 | Training Set | 0.05219575 | 0.9317544 | 0.03353768 |
| Testing Set | 0.10141311 | 0.7818169 | 0.06360183 |
| XG Boost | 70-30 | Training Set | 0.06856223 | 0.8656509 | 0.03880017 |
| Testing Set | 0.08370564 | 0.8037405 | 0.04663333 |
| 80-20 | Training Set | 0.06627407 | 0.8749223 | 0.03744300 |
| Testing Set | 0.09847384 | **0.7852130** | 0.05784156 |
| Gradient Boosting | 70-30 | Training Set | 0.06688688 | 0.8722068 | 0.03771209 |
| Testing Set | 0.08331763 | 0.8058273 | 0.04593434 |
| 80-20 | Training Set | 0.06632627 | 0.8747767 | 0.03781746 |
| Testing Set | 0.09902834 | 0.7786041 | 0.05805065 |
| SVM  Regressor | 70-30 | Training Set | 0.06800474 | 0.8686853 | 0.03325114 |
| Testing Set | 0.08328600 | 0.8096834 | 0.04268856 |
| 80-20 | Training Set | 0.06807194 | 0.8688119 | 0.03336189 |
| Testing Set | 0.09972658 | 0.7827709 | 0.05623169 |
|  |  |  |  |  |  |  |
| 9587 | Random Forest | 70-30 | Training Set | 0.05157841 | 0.9318347 | 0.03242416 |
|  | Testing Set | 0.09908885 | 0.7962776 | 0.06215094 |
| 80-20 | Training Set | 0.05174449 | 0.9314813 | 0.03254364 |
|  | Testing Set | 0.10510476 | 0.7741989 | 0.06636156 |
| XG Boost | 70-30 | Training Set | 0.06852550 | 0.8635188 | 0.03777131 |
|  | Testing Set | 0.09411924 | 0.8037324 | 0.05435846 |
| 80-20 | Training Set | 0.06551298 | 0.8755104 | 0.03641165 |
|  | Testing Set | 0.09880379 | **0.7858113** | 0.05711739 |
| Gradient Boosting | 70-30 | Training Set | 0.06903272 | 0.8623004 | 0.03820331 |
|  | Testing Set | 0.09463114 | 0.8060675 | 0.05562951 |
| 80-20 | Training Set | 0.06596047 | 0.8738091 | 0.03679935 |
|  | Testing Set | 0.09864245 | 0.7838222 | 0.05619292 |
| SVM  Regressor | 70-30 | Training Set | 0.06691112 | 0.8704024 | 0.03337734 |
|  | Testing Set | 0.09472633 | 0.8078655 | 0.05257949 |
| 80-20 | Training Set | 0.06686211 | 0.8707424 | 0.03303151 |
|  | Testing Set | 0.10122340 | 0.7840808 | 0.05695471 |

Fig: Model Summary

R-squared will be used to compare the prediction accuracy of different models. R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.

Insights from Model Summary:

* Utilizing the same seed i.e. 3238, different train-test size was tested i.e. of 70-30 and 80-20. It did not result any significant change in individual model performances. XG Boost has the highest out sample R-square for 80-20 split while Random Forest has the highest out sample R-square in 70-30 split
* Random Forest shows a high in-sample R-square value of ~ 93% which indicates that Random Forest is overfitting the data. Consequently, it’s out-sample R-square value is significantly low from the in-sample for both 70-30 and 80-20 split
* With the seed value of 9587, for 80-20 split, XG Boost outperforms the other models with highest R-square value. And, for 70-30 split, SVM regressor outperforms the other models with slightly higher R-square value than XGBoost.

Extreme Gradient Boosting (XG Boost) consistently has the highest out-sample R-square value in both cases, i.e. with seed 3238 and 9587. This can be stated to the fact, that XG Boosting uses a more regularized model formalization to control over-fitting, which gives it better performance.

Thus, XG Boost represents the best model with highest accuracy measured in terms of out-sample R-squared value of ~ 79% (0.786).

1. **Conclusion**

Different supervised learning algorithms were utilized to accurately predict the target regressor i.e. charges. Feasibility study was done using Linear Regression. This pointed about the interesting interaction effects existing in data. Various models such as Random Forest, Gradient Boosting, etc were applied and below graph was obtained.

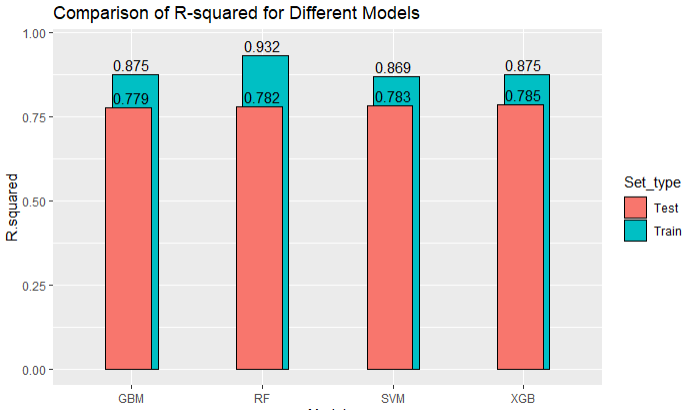


Fig: Comparison of R Squared for different models

The figure shows that XGBoost has the highest out-sample performance due to the highest R-square. Other models like Gradient Boosting was also quite close with almost similar out-of-sample R-squared value. Insurance charges can be best predicted using XG Boost Model using the predictors in the model formula. For a business perspective, insurance charges can be predicted using information such as age, bmi30: smokeryes, etc for individual customers. This could also derive customer loyalty as the insurance charges suits their income and category.