**Interim Report of**

**Health Care Project**

**Submitted By**

**Batch: August 2021 (LVC)**

**Group Members**

1. **Ravindra Tabde**
2. **Santosh Ingle**

**Research Supervisor**

**Ms. Indumathi Taneja**

Contents

[**1. Introduction**](#_Toc40201628) **4**

[**2. Problem Statement, Scope and Objective**](#_Toc40201629) **4**

[**3. Data Source and Description**](#_Toc40201630) **5**

[**4. Exploratory Data Analysis**](#_Toc40201631) **7**

[**5. Data Cleaning and Preprocessing**](#_Toc40201632) **12**

[**6. Modelling Approach**](#_Toc40201633) **13**

[**7. Model validation**](#_Toc40201634) **17**

[**8. Bussiness Insights and recommendations from EDA**](#_Toc40201635) **19**

[**9. References and Bibliography**](#_Toc40201636) **20**

**10. Appendix ………………………………………………………………………………………….... 20**

Table

[**1. Data dictionary**](#_Toc40201628) **5**

[**2. Various Models performance matrix**](#_Toc40201629) **18**

1. **Introduction**

We are on a planet full of threats and uncertainty. People, households, companies, properties, and property are exposed to different risk forms. And the risk levels can vary. These dangers contain the risk of death, health, and property loss or assets. Life and wellbeing are the greatest parts of people's lives. but, risks cannot usually be avoided, so the world of finance has developed numerous products to shield individuals and organizations from these risks by using Financial capital to reimburse them. Insurance is, therefore, a policy that decreases or removes loss costs incurred by various risks.

According to World Bank data India are spending up to 3.01 % of GDP. [1]. Health insurance is one of the most significant investment an individual makes every year.

According to the Department of Health and Family Welfare has gone up from Rs 2508 crore in FY2021-22 to Rs 5632 crore for FY2022-23, a more than 100% increase to building strong foundations for health system.[2].

1. **Problem Statement, Scope and Objective**

**2.1. Problem Statement:**

We all know that Health care is very important domain in the market. It is directly linked with the life of the individual; hence we have to be always be proactive in this particular domain. Money plays a major role in this domain, because sometime treatment becomes super costly and if any individual is not covered under the insurance, then it will become a pretty tough financial situation for that individual. The companies in the medical insurance also want to reduce their risk by optimizing the insurance cost, because we all know a healthy body is in the hand of the individual only. If individual eat healthy and do proper exercise the chance of getting ill is drastically reduced.

**2.2. Need of the study/project**

The objective of this exercise is to build a model, using data that provide the optimum insurance cost for an individual. We have to use the health and habit related parameters for the estimated cost of insurance.

**2.3. Understanding Business/Social Opportunity**

Based on the study it could be very advantageous to estimate the insurance cost of candidate, to go for plans. The insurance cost predictors may serve as a guide to related stakeholders for present as well as future.

1. **Data Source and Description:**

|  |  |  |
| --- | --- | --- |
| Variable Name | Description | Data Type |
| Applicant id | Applicant unique ID | int64 |
| Years of insurance with us | Since how many years customer is taking policy from the same company only | int64 |
| Regular check-up last year | Number of times customers has done the regular health check-up in last one year | int64 |
| Adventure sports | Customer is involved with adventure sports like climbing, diving etc | int64 |
| Occupation | Occupation of the customer | object |
| Visited doctor last 1 year | Number of times customer has visited doctor in last one year | int64 |
| Cholesterol level | Cholesterol level of the customers while applying for insurance | object |
| Daily avg steps | Avarage daily steps walked by customers | int64 |
| Age | Age of the customer | int64 |
| Heart decs history | Any past heart diceases | int64 |
| Other major decs history | Any past major diceases apart from heart like any operation | int64 |
| Gender | Gender of the customer | object |
| avg\_glucose\_level | Average glucose level of the customer while applying the insurance | int64 |
| Bmi | BMI of the customer while applying the insurance | float64 |
| smoking\_status | Smoking status of the customer | object |
| Year\_last\_admitted | When customer have been admitted in the hospital last time | float64 |
| Location | Location of the hospital | object |
| Weight | Weight of the customer | Int64 |
| covered\_by\_any\_other\_company | Customer is covered from any other insurance company | object |
| Alcohol | Alcohol consumption status of the customer | object |
| exercise | Regular exercise status of the customer | object |
| weight\_change\_in\_last\_one\_year | How much variation has been seen in the weight of the customer in last year | int64 |
| fat\_percentage | Fat percentage of the customer while applying the insurance | int64 |
| insurance\_cost | insurance\_cost: Total Insurance cost | int64 |

The dataset consists of 25000 rows and 24 features. The features are as in below table:

Table 1 Data Dictionary

3.1 Understanding how data was collected in terms of the time, frequency and methodology

There might chance that data is collected by a market research firm from medical firm across top Indian cities. Variables related to health are enlisted and recorded,

Insurance cost of the data collected ranges from 2468 to 67870.in the Dataset applicant details is provided with respect to various parameter related to habits and health like weight, bmi, smoking , exercise and drinking habits, medical history etc. are recorded.

3.2. Visual inspection of data1 (rows, columns, descriptive details)

* 14 Variables are in integer datatype, 8 Variables are in Object data type and rest 2 Variables are in float data type
* Some column names to be ambiguous. But from description the variable definition its more clear.
* Some of the variables with numerical values are acting like categorical variables like heart\_decs\_history, other\_mojor\_decs\_history, adventure sports etc.
* Variable like bmi and Year\_last\_admitted has have missing values which needs to be treated
* Range of some variables is very large, indicating presence of possible outliers, which can be confirmed through boxplots.

1 Refer Annexure no. 1

**4. Exploratory Data Analysis**

4.1 Univariate analysis2

We can see, there are features which have outliers. So, we might need to treat those before building model (Annexure 2, Fig no. 1)

1. **Applicant\_id**: we have total 25000 applicant in data set. With no duplicate records.
2. **Years of insurance with us**: Insurance with us ranging from 0 – 8 years. We have treat this variable as category as it has only 9 unique values which is discreet. (Annexure 3, fig no.14)
3. **Regular check-up last year:** This variables has value ranges from 0-5 year in this case also we can treat this variable as categorical as it has 6 unique values. The applicant who have not check-up last year is about 60%.(Annexure 3, fig no.15)
4. **Adventure sports:** It is Boolean datatype variable as it has only 0 and 1 entries. Most of the people do not take part in adventures sport. Only 8 % people take part in adventures sports. (Annexure 3, fig no.2).
5. **Occupation:** There are 3 difference subcategories in this variable student, business and Salaried. Most of the applicant comes under the student and business category. (Annexure 3, fig no.2)
6. **Visited doctor last 1 year:** There are 2,3 and 4 doctor visits are most common for applicant. We can treat this variable as categorical as it has only 12 unique values. (Annexure 3, fig no.17)
7. **Cholesterol level:** Almost 69 % of customer has cholesterol level in between 125 -175. 8.2 % of customer have very high cholesterol level that is in between 225-250.
8. **Daily avg steps:** it is ranging from 2400 to 11255 and distribution is slight right-skewed. There are outlier in the dataset for this variable only 1512 records has avg step more than 7000. We can treat this variable prior to model building. (Annexure 2, fig no.1)
9. **Age**: Age values ranging from 16 to 74 years. Mean and median for age is almost equal to 45. The distribution looks likes normal without having outlier in the data. (Annexure 4, fig no.20)
10. **Heart decs history:** Majority of applicant do not have heart decs history (~ 95%).

2 Kindly refer annexure no 2 , 3 and 4

1. **Other major decs history**: Majority of applicant do not have other major decs history (~ 90%).(Annexure 3, fig no.6)
2. **Gender:** According to dataset there are more male customer than female customer.
3. **Avg\_glucose\_level:** Avg glucose values ranging from 57 to 277 years. Mean and median for age is almost equal to 168. The distribution looks likes normal without having outlier in the data. (Annexure 4, fig no.21)
4. **bmi** :There are 4 % missing values in the data for bmi variable.It is ranging from 12.3 to 100.6. bmi variable distribution is right skewed having outlier at both sides. (Annexure 4, fig no.22)
5. **Smoking\_status:** majority of customer never smoke and almost 30% customer whose status is unknown. As the domain perspective this variable must have higher weightage in the insurance cost.
6. **Year\_last\_admitted:** more than 47 % of data is missing in this variable. But domin perspective this variable is more significant so we can do feature engineering and come up with new feature for further analysis. Year admitted history from 1990 to 2018 is recorded in dataset.
7. **Location:** in this 15 different location information given for respective individuals .almost all city have equal applicant that is 6-7 %. Bangalore is the location which have maximum number of applicant 1742.
8. **Weight:** weight values ranging from 52 to 96 kg. Mean and median for weight is almost equal to 72 kg. The distribution looks likes normal without having outlier in the data. (Annexure 2, fig no.1 and Annexure 4, fig no.23 )
9. **covered\_by\_any\_other\_company:** Majority of individuals do not have insurance covered with any other company (~70%).
10. **Alcohol:** Almost 11 percent of individuals have daily habit of drinking whereas 34 % do not drink alcohol.(Annexure 3, fig no.11)
11. **Exercise:** majority of individuals do moderate exercise that is almost 57 %.
12. **weight\_change\_in\_last\_one\_year**: In this variable we have 6 unique values due to which we can treat this variable as categorical instead continuous. Majority of individuals have weight change between 2-3 kg in last one year.
13. **fat percentage**: it is ranging from 11 to 42 and distribution is slight left-skewed. There are no outlier in the dataset for this variable. Building. (Annexure 2, fig no.1)
14. **insurance cost:** This is the our target variable. Insurance cost is ranging from 2468 to 67870. This variables do not have outlier and mean and median values are almost equal to 27148.

4.2Bivariate analysis I:

From Heatmap3

* Heat map of continuous variables indicate there is strong correlation between **insurance cost** and **weight** variable. (Annexure 5, fig no.25)
* However, heat map is a better tool to understand the extent of correlation between the variables. It showed that there is **no strong positive or negative correlation between** any of the other variables except insurance cost and weight.
* Heat map shows that this data is not fit for linear regression analysis as it does not hold very first assumption of linear regression that independent variable must show linear relationship with the dependent or target variable.
* We can explore the analysis of every independent variable with respect to the target variable and try to find out the insights based on it.

4.3Bivariate analysis II:

1. **Years of insurance with us Vs. Insurance cost:**

No clear relationship with insurance cost or any other feature. 9 unique values so can be converted to Categorical Variable.

1. **Regular check-up last year :**

It shows slightly negative relationship with target variable. But which is not that much significant so we can convert it into categorical variable.

1. **Adventure sports**: The individual take part in adventures sport has has high average insurance cost than on playing individuals. (Annexure 6, fig no.33)
2. **Occupation**: Average insurance cost for business person is higher than the student and salaried individuals. (Annexure 6, fig no.26)
3. **Visited doctor last 1 year:** No clear relationship with insurance cost or any other feature. 12 unique values so can be converted to Categorical Variable.

3 Kindly refer attached python notebook file Ravindra\_Tabde\_15052022

1. **Cholesterol level :** average Insurance cost for cholesterol level 200-225 is higher than the other levels**.** (Annexure 6, fig no.31) According to dataset surprisingly avg insurance cost for 125-150 is more than that of highest cholesterol level 225-250.
2. **Daily avg steps:** No clear relationship with insurance cost and Daily avg steps.

(Annexure 7, fig no.36)

1. **Age:** In the general case domain perspective as age increases insurance cost also increases. But in case of our dataset which is not true. There is no relationship between age and insurance cost.
2. **Heart decs history:** Insurance cost range is quite high for individual having heart desc history (approx. 26400 to 27800.) (Annexure 6, fig no.34)
3. **Other major decs history:** According to data Individual having other desc history have less average insurance cost than the individual do not have any desc history. In this case we need to explore the combined other variable effect on the insurance cost. (Annexure 6, fig no.35)
4. **Gender:** average insurance cost for female is greater than the male and it also has wide range 26900 to 27500 approx. (Annexure 6, fig no.27)
5. **avg\_glucose\_level:** in ideal case if avg glucose level is very high or very low that individual have high insurance cost. But in our data there is no relation between the glucose level and insurance cost**.** (Annexure 7, fig no.36)
6. **Bmi**: bmi is the index of body shape. Whether the individual is underweight, overweight or obese. According to our data bmi do not have direct linear relationship with the insurance cost. (Annexure 7, fig no.36)
7. **Smoking status:** In the healthcare domain the smoking is the critical factor and more direct affecting factor on insurance cost.According to our dataset average insurance cost is high for smoking status unknown category individuals. Followed by never smoke individual. We cannot make judgment in present smoking status variable only whether the insurance cost is high or low. But we need to examine the combine effect in presence of other variable. (Annexure 6, fig no.28)
8. **Location**: Average insurance cost for Mangalore and Guwahati is high as compare to other city. Whereas for Bangalore and Ahmedabad it low.
9. **Weight**: weight is the highly correlating factor with the insurance cost according to our data. There is high positive correlation between the weight and insurance cost. (Annexure 7, fig no.36)
10. **covered\_by\_any\_other\_company**: Average insurance cost for individual covered by any other company is higher than not covered individuals. (Annexure 6, fig no.32)
11. **Alcohol:** According to our dataset average insurance cost is high for non-drinker individuals. Followed by rare drinker individual. Surprisingly average insurance cost for daily drinker is low than other individuals. We need to examine the combine effect in presence of other variable. (Annexure 6, fig no.37)
12. **Exercise**: According to our dataset average insurance cost is low for individual who can’t do any kind of exercise. We need to examine the combine effect in presence of other variable. (Annexure 6, fig no.30)
13. **weight\_change\_in\_last\_one\_year:** data shows that the as weight change in last one year increases the insurance cost decreases. But this relationship is not significant to that extend so we can convert this variable into categorical variable for further model building (Annexure 7, fig no.36)
14. **fat\_percentage:** No clear relationship with insurance cost and fat percentage.(Annexure 7, fig no.36

3 Kindly refer attached python notebook file Ravindra\_Tabde\_15052022

**5. Data Cleaning and Preprocessing4**

**5.1 Missing value treatment**

* Here only two features have missing values that is year\_last\_addmitted and bmi.
* For our business point of view these two feature is very important.
* Bmi feature have almost 4 % missing values whereas year\_last\_addmitted almost have 48 % missing values.
* If we observed values from year\_last\_addmitted feature it has NaN values for year so we can assume that this person never been admitted before. We can do feature engineering here and create new addmitted\_status yes or No. There might be chance of person with higher age and admitted status no.
* For bmi feature we can impute the missing values base on the gender.
* There are 733 Male and 257 female record for which bmi value is missing
* There are difference in the median value of bmi across gender.
* We can impute the missing values BMI for Male and Female with the median value of male and female bmi respectively.

**5.2. Outlier treatment**

* Though many outliers are observed in some continuous variables, they cannot be dropped from the data straightaway.
* Treating outliers, we need to be cautious, we must not remove them without proper analysis as some variables may have outliers and may represent some meaning to the data. We are treating the outliers with IQR\*1.5
* We got 589 records which are oulier in the bmi variable.
* This might be the high insurance cost customer, so we have to keep this customer by simply treating the oulier not dropping it.
* Data have 921 records as outliers for daily avg step variable, lets also treat these outlier point
* After treating outliers of daily\_avg\_steps, data and bmi the distribution of both variable looks nicely distributed.

**5.3. Removal of Unwanted Variable and Renaming of Features**

* Applicant\_id is serial number and has a unique value for each row. Since it won’t help us for further analysis, hence we are dropping this column.
* There are few variable having wrong name by meaning and by conventions. We can rename it before the model building.

4 Kindly refer attached python notebook file Ravindra\_Tabde\_15052022

**5.4. Encoding of variables**

* Categorical variables require special attention in regression analysis because, unlike dichotomous or continuous variables, they cannot by entered into the regression equation just as they are. Instead, they need to be recoded into a series of variables which can then be entered into the regression model.
* For model building dummies variables are created for categorical variables: adventure\_sports, Occupation, cholesterol\_level,heart\_decs\_history, other\_major\_decs\_history, Gender, smoking status, Location, covered\_by\_any\_other\_company, Alcohol, exercise, years\_of\_insurance\_with\_us,visited\_doctor\_last\_1\_year, regular\_checkup\_lasy\_year, admitted status.
* Now we have total 65 variable and 25000 records for model building after encoding.

**6. Modelling Approach (Predictive approach)**

* After data processing data was split in test and train with 70:30 ratio with random state of 1.
* Since the target variable is insurance price which is continuous integer data type, regression models were applied on the data for predicating the insurance cost.

**6.1 Regression models used for prediction** **5**

6.1.1 Linear Regression –

Attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable.

**The linear regression model performed with scores 0.944 & 0.945 in training data set and validation data set respectively.**

6.1.2 Linear Regression (Lasso)

**Lasso regression** is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

5 Kindly refer attached python notebook file Ravindra\_Tabde\_15052022

**The lasso linear regression model performed with scores 0.944 & 0.945 in training data set and validation data set respectively.** The coefficient’s of few variable in lasso model is almost '0', signifying that the variable with '0' coefficient can be dropped.

6.1.3 Ridge Regression

It is a technique for analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors.

**The lasso linear regression model performed with scores 0.944 & .945 in training data set and validation data set respectively**.

The coefficeints of variables in ridge model are all non-zero, indicating that none of the variables can be dropped.

6.1.4 KNN Regression:

KNN regression is a non-parametric method that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighbourhood

**The KNN regressor model performed with scores 1.00 & 0.58 in training data set and validation data set respectively. It is over fitting in the training set and performed poor in to the validation**

6.1.5 Decision Tree Regression

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes

DT1 - DecisionTreeRegressor()

**The Decision tree regressor model performed with scores 1.00 & 0.91 in training data set and validation data set respectively**. It is overfitting in the training set and performed poor in to the validation.

DT2 - DecisionTreeRegressor(max\_depth=10, min\_samples\_leaf=5)

The Decision tree regressor model with modified parameter performed with scores 0.96.4 & **0.945 in training data set and validation data set respectively**. Above decision tree model with modified parameter has better performed on the training set and validation set compared to initial decision tree model. This DT models have overall decision tree performed well than linear regression models. As is also have low RMSE value than all other models. But still 3334.9 RMSE is large error value.

**6.2 Ensemble modelling6**

Ensemble methods are techniques that create multiple models and then combine them to produce improved results. Ensemble methods usually produces more accurate solutions than a single model would. This has been the case in a number of machine learning competitions, where the winning solutions used ensemble methods

**6.2.1 Boosting and Bagging**

Boosting is an iterative technique which adjusts the weight of an observation based on the last classification. Bagging is a way to decrease the variance in the prediction by generating additional data for training from dataset using combinations with repetitions to produce multi-sets of the original data

GB1: GradientBoostingRegressor(n\_estimators = 200, learning\_rate = 0.1, random\_state=22) **The Gradient boosting performed with scores 0.958 & 0.956 in training data set and validation data set respectively**. Gradient boosting model has provided good scores in both training and validation sets. Also RMSE value is just below 2995.

BGG1: BaggingRegressor (n\_estimators=50, oob\_score= True, random state=14)

**The Bagging model performed with scores 0.99 & 0.95 in training data set and validation data set respectively**. It is overfitting in the training set and performed better in to the validation. need to analyse further by hyper tuning. But RMSE value is more than boosting model.

6 Kindly refer attached python notebook file Ravindra\_Tabde\_15052022

**6.2.2 Random Forest**

Random forest is an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees.

**The Random Forest model performed with scores 0.99 & 0.95 in training data set and validation data set respectively.** It is over fitting in the training set and performed better in to the validation.

**6.3 Steps to improve model performance**

* Few models showed over fitting in train data, which can be improved by feeding more data but already we splatted 80 percent of data. Using regularization, it can be solved but we haven’t used here.
* Iterative process for Decision tree with modified parameter was done and performance was changed in it and over fitting also got solved
* Similar to Decision tree we have done iterative process for Boosting and bagging models Gid search parameter is also a way to fine best parameter

# 7. Model validation

Regression model’s validation is the process of deciding whether the numerical results quantifying hypothesized relationships between variables, obtained from regression analysis, are acceptable as descriptions of the data. The validation process can involve analyzing the goodness of fit of the regression, analyzing whether the regression residuals are random, and checking whether the model's predictive performance deteriorates substantially when applied to data that were not used in model estimation.

Different models, which performed relatively better, compared to other models are identified. Model evaluation was done based on the R2, RMSE values, MSE value and MAE value

**1. R-squared**

It is a goodness-of-fit measure for linear regression models. This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively. R- squared measures the strength of the relationship between your model and the dependent variable on a convenient 0 – 100% scale

**2. RMSE**

The RMSE is the square root of the variance of the residuals. It indicates the absolute fit of the model to the data–how close the observed data points are to the model’s predicted values. Whereas R-squared is a relative measure of fit, RMSE is an absolute measure of fit. As the square root of a variance, RMSE can be interpreted as the standard deviation of the unexplained variance, and has the useful property of being in the same units as the response variable. Lower values of RMSE indicate better fit. RMSE is a good measure of how accurately the model predicts the response, and it is the most important criterion for fit if the main purpose of the model is prediction

**3. MSE**

The mean squared error tells you how close a regression line is to a set of points. It does this by taking the distances from the points to the regression line (these distances are the “errors”) and squaring them. The squaring is necessary to remove any negative signs. It also gives more weight to larger differences. It’s called the mean squared error as you’re finding the average of a set of errors

**4. MAE**

Mean Absolute Error (MAE) is another loss function used for regression models. MAE is the sum of absolute differences between our target and predicted variables. So it measures the average magnitude of errors in a set of predictions, without considering their directions.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Method  Linear Reg Model1 | Val Score | RMSE\_vl | MSE\_vl | MAE\_vl | Train\_Score | RMSE\_Train | MSE\_Train | MAE\_Train |
| 0.9453 | 3349 | 1.12E+07 | 2726 | 0.9449 | 3362 | 1.13E+07 | 2712 |
| Linear-Reg  Lasso1 | 0.9454 | 3347 | 1.12E+07 | 2724 | 0.9449 | 3363 | 1.13E+07 | 2713 |
| Linear-Reg  Ridge1 | 0.9453 | 3349 | 1.12E+07 | 2726 | 0.9449 | 3362 | 1.13E+07 | 2712 |
| knn1 | 0.5886 | 9188 | 8.44E+07 | 7257 | 1.0000 | 0000 | 0000 | 0000 |
| DT1 | 0.9104 | 4286 | 2.92E+07 | 3328 | 1.0000 | 0000 | 0000 | 0000 |
| DT2 | 0.9458 | 3335 | 2.38E+07 | 2649 | 0.9642 | 2708 | 0.73E+07 | 2135 |
| GB1 | 0.9563 | 2994 | 0.89E+07 | 2418 | 0.9581 | 2933 | 0.86E+07 | 2363 |
| BGG1 | 0.9512 | 3139 | 0.98E+07 | 2528 | 0.9929 | 1202 | 0.14E+07 | 944 |
| RF1 | 0.9525 | 3122 | 0.97E+07 | 2515 | 0.9933 | 1173 | 0.14E+07 | 929 |

Table 2 Various Models performance matrix

**7.1 Model evaluation interpretation**

1. Model score (R2) for validation set is highest for gradient boosting, random forest shows the ensemble techniques has given better fit in validation set. Whereas the linear regression has given score of 0.94 slightly less than the ensemble techniques. Poor score was in KNN regressor
2. Model score (R2) for train set is highest BGG1, RF,KNN and DT1 the value is above 0.95 which shows data is overfitting in train set which need to be regularise. The preferable score for train for GB! And Random forest regressor.
3. For Validation set RMSE is preferable in ensemble models boosting and random forest models on other side it is high in KNN, DT1 and DT2.
4. For Validation set similar to RMSE, MSE and MAE is preferable in ensemble models and Random forest regressor models on other side it is high in KNN, DT1 and DT2
5. For Validation set RMSE is preferable in ensemble models and non-preferable linear regression models due to high value
6. For Validation set similar to RMSE, MSE and MAE is preferable in ensemble models and non- preferable for linear regression models due to high value
7. In short ensemble models gradient boosting and Random forest have performed relative better among all the models used. Since ensemble models gradient boosting regression is perform better than other model so we will **select models gradient boosting regression**.

**8. Business Insights and recommendations from EDA**

* Weight is the one of the most significant features in the data which has highest influence on insurance cost, insurance provider always consider this feature at most important while calculating insurance cost.
* Based on the model the insurance cost predictors having high importance or the top key features to consider for insurance costing are weight, age, average glucose level , daily avg\_steps,and bmi 5 variables preserve 97.11% of variance of the data.(Annexure 9 fig no.39)
* Average insurance cost (28,849) is higher for daily drinker, without exercise and smoking status smoke. Which is 1531 more than the individuals who not drink, not smoke and do moderate exercise.
* We need to examine important variable like alcohol, smoking status whether there is data inadequate or other measures affecting this variable.

**9. References and Bibliography**

1. World Helath Organization Global Health Expenditure database “Current health expenditure

(% of GDP)”, last updated 02 May 2022, View at: [Publisher Site](https://data.worldbank.org/indicator/SH.XPD.CHEX.GD.ZS?locations=IN)

1. Department of Health and Family Welfare India, “Union Budget 2022 & the big health push”

View at: [Publisher Site](https://economictimes.indiatimes.com/news/economy/policy/view-budget-2022-and-the-big-health-push/articleshow/89572983.cms?utm_source=contentofinterest&utm_medium=text&utm_campaign=cppst)

1. David S. Younger, MD, MPH, MS “Health Care in India” [Volume 34, Issue 4](https://www.sciencedirect.com/journal/neurologic-clinics/vol/34/issue/4), November 2016, Pages 1103-1114 , View at: [Publisher Site](https://www.sciencedirect.com/sdfe/pdf/download/eid/1-s2.0-S0733861916300275/first-page-pdf)

**10. Appendix**

Separate Annexure document provided in the .ppt format.

[ *Ravindra\_Tabde\_Annexure\_health\_care\_capstone\_note.pptx*].