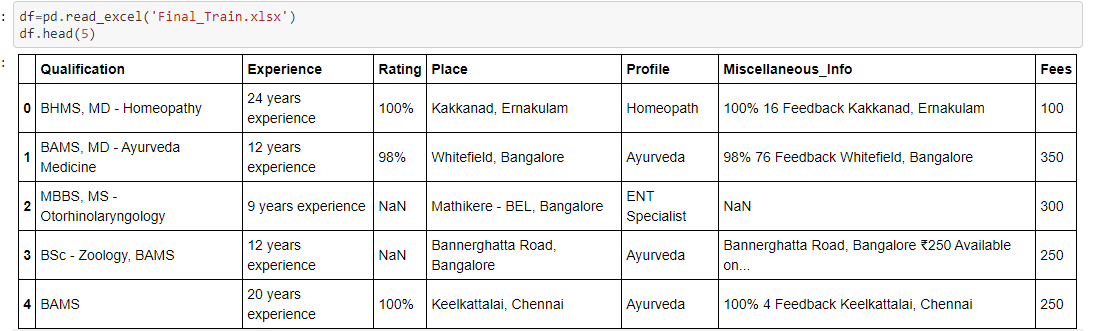
# **Doctor's Consultation Fee analysis**

# **Problem statement:**

# **Before going to any doctor, we always wonder what the doctor consultancy fee would be and what amount they are going to charge us. I am young and haven’t been to any doctors for any kind of consultation fortunately till now. But for my parents I have been to few clinics and I always do my research about what kind of doctor I am going to, what are their qualification, what are their specialisation, how many years of experience do they carry etc. All this information is needed and according to all this criteria doctors charges their fees. In this data analysis we will be building the model where we will be predicting the doctor consultancy fees.**

# 

In todays blog we will use the dataset about doctor consultation and try to look at analytical point of view. Let’s have the look at the dataset.



Features:

Qualification: Qualification and degrees held by the doctor

Experience: Experience of the doctor in number of years

Rating: Rating given by patients

Profile: Type of the doctor

Miscellaeous\_Info: Extra information about the doctor

Fees: Fees charged by the doctor

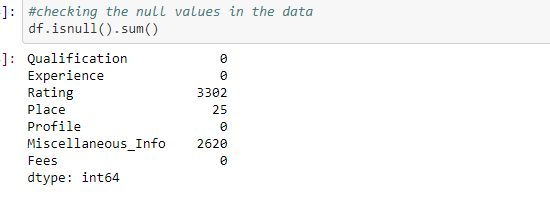
Place: Area and the city where the doctor is located.

Pre processing:

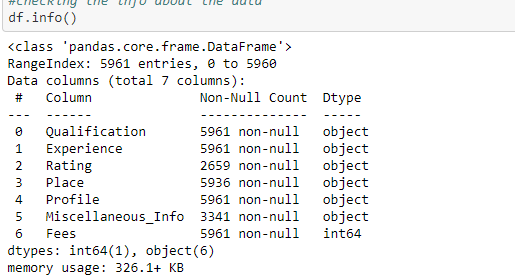
The first step will be importing all the necessary libraries in python and loading the csv files which contains all the data about the doctor’s consultation fee.

After looking at the dataset it looks like it requires a bit of cleaning for example,

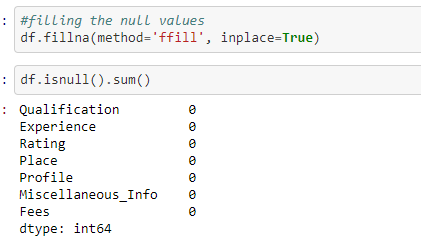
1)there are NaN values present in the dataset



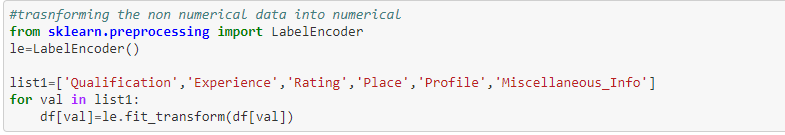
2) all the columns are object type except for the “Fees” column.



Filling the NaN values.



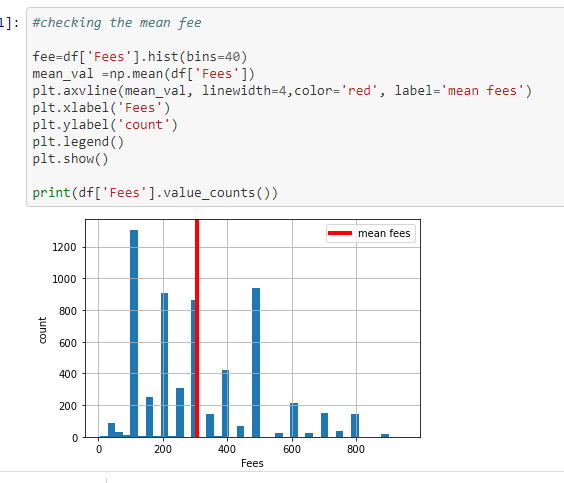
Encoding the object values



EDA process:

After cleaning the data, we are ready for the visualization part of the data where we can better understand our data and learn more about it by using MATPLOTLIB and SEABORN.

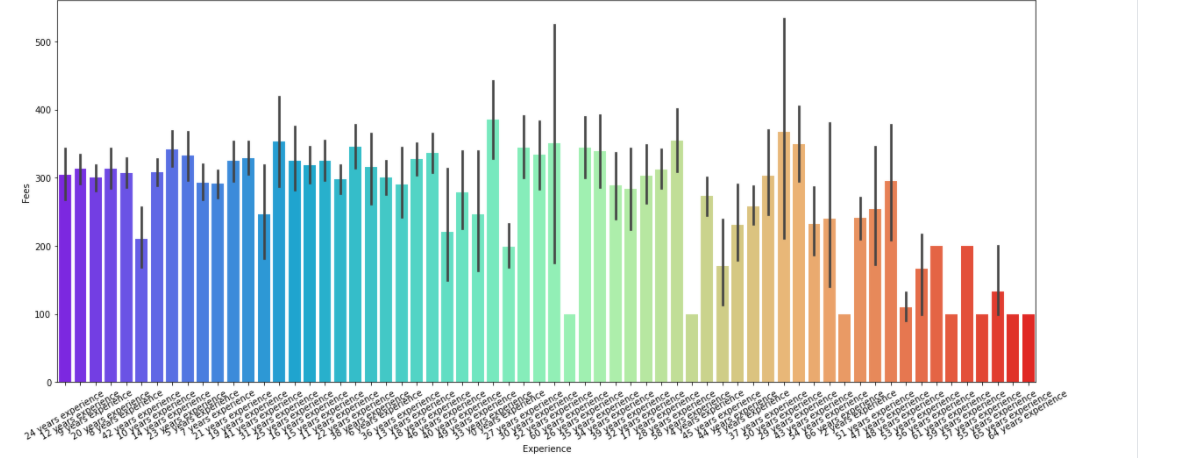
Mean of consultation fees:



First, here is a general histogram of doctor consultation fees. We have plotted the mean fees indicated by the red line. As we can clearly see this takes on standard, normally distributed shape, with an average fee ranging from 0 to 800 where 0 being the minimum and 800 being the maximum.

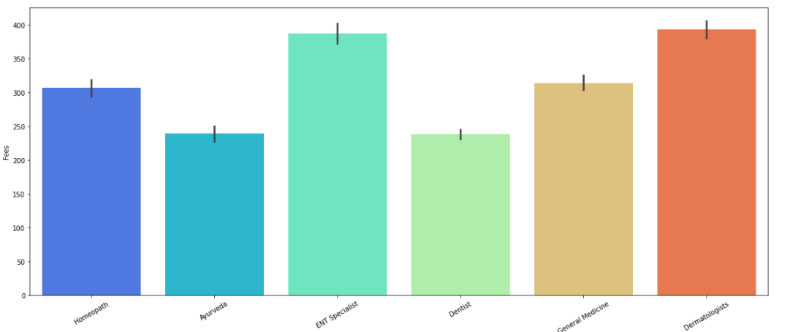
But we would want to explore the relationship of doctor consultation fee with other columns to have a better understanding of what factors affects the fees so let’s plot a bar plot and compare the columns with fees.

Experience Vs Fees:



This is more interesting. we can see from the boxplot which has explored the relationship of fees with experience and how the fee varies depending upon level of experience. However, there is not much significant hike or decrease in fees with level of experience. We can see from the above bar plot that fees are low with some doctors although having 50 plus experiences and fees being high for doctors having 10 plus experience so we can conclude that experience isn’t contributing much to the doctor’s consultation fees.

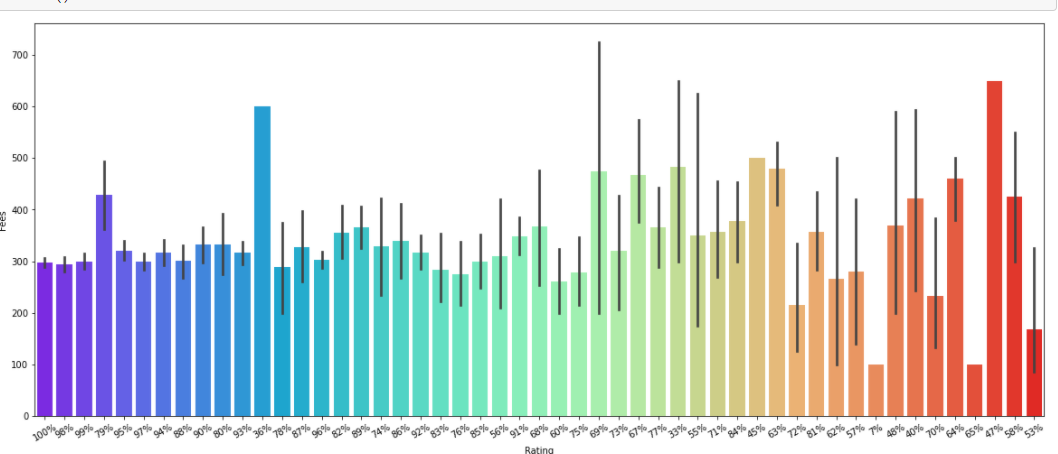
Profile vs fees:



As we can see there is significant differences in fees across profile. There are five six profile which are -Hemopathy, Ayurveda, ENT Specialist, Dentist, General Medicine and dermatologists.

Dentist and Ayurveda seems to have cheapest consultation fees closely followed by homeopathy. Two profile with highest consultation fees are dermatologist and ENT Specialist.

Rating vs fees:

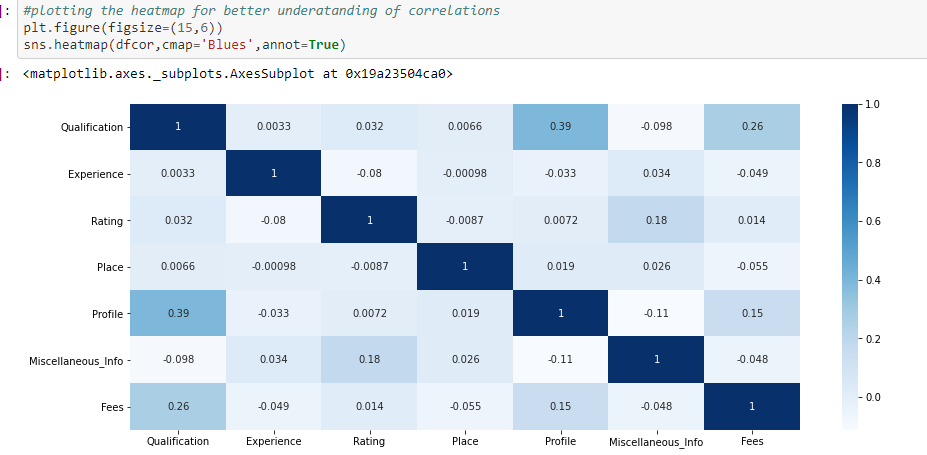


This is interesting.as we can see from this plot which depicts the changes of fees with change in ratings and we can say that it is very dynamic.at 100 % percent rating fees is minimum while at the same time at 47% rating fess is maximum.

Now that the visualization gives us the better idea about what is the relationship of doctor’s fee with other columns, lets change course and ask more fundamental question: what is the correlation between dependent variables and independent variable.

**Correlation:**

its important to understand the correlations with the target variable which is doctors’ fees. The correlation plot and table is manly to show to what extent the variables are related to each other. I have plotted a heatmap where the darker colour will represent higher correlation with the target variable and the lighter colour will represent the low correlation with the target variable.

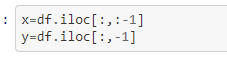


I have plotted the heat map from where we can analysis the correlations. As we can see the job profile of the doctor and qualification of the doctor has highest correlation with the amount of fee charged. Experience and ratings have a negative correlation with the target variable.

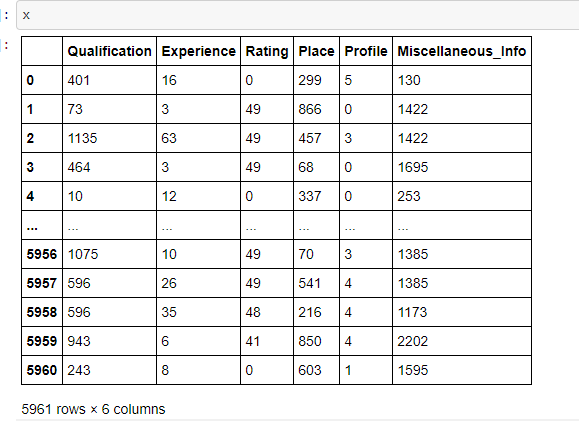
**EDA concluding remarks:**

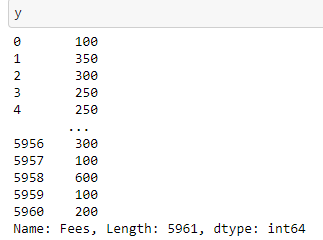
Now, combining the information from the above graphs we would expect a hike in doctors consultation fees if both the job profile and qualification are at the highest standard and that is exactly what we see. Clearly there is a huge difference in fees when the doctor belongs to profile such as ENT specialist then Dentist. Since qualification has highest correlation with Fees it will directly impact the amount of fee charged. Also, we have other features but with negative correlation or very low correlations Next step will be scaling the data and building the model where we will test various models to find the best possible outcomes.

**Splitting the data:**

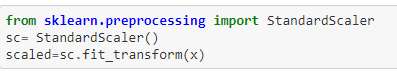
****

We are splitting the variable using the” iloc function” into x which is our dependent variable and into y which is our target variable.



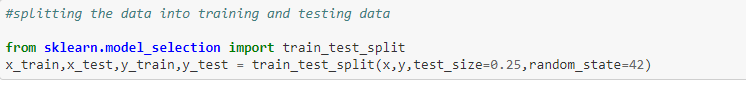
****

after splitting the data we can scale the x variable using the standard scaler.



We are importing StandardScaler from the scikit learn and scaling all the x variables so we can use this for model building which is our next step but before that we are splitting the data into training and testing.

**Splitting the data:**



We have split the data into train and test with test size of .25 and also provide with the random state of 42.now we can move towards our model building stage and find the accuracy of different models.

**Model building:**

This is one of the most important stage in analysing the data. We will be importing various libraries from scikit learn which will help us in model building. Since it’s a linear regression problem we will be using linear regression algorithms

Importing all the library



So we have imported the necessary libraries to predict the best possible outcomes.frist lets understand each machine learning algorithm one by one, how it function and understand how will it help in predicting the doctor consultation fees.

**Linear Regression:**

Linear regression is one of the most popular and best understood algorithms in the machine learning landscape. Since regression task belong to the most common machine learning problems in supervised learning, every machine learning engineer should have a thorough understanding of how it works. This blogpost covers how the linear regression works, so with the help of linear regression and its various tools and techniques we are trying to find the best accuracy.

It is used to estimate real values (cost of houses, number of calls, total sales etc.) based on continuous variable(s). Here, we establish relationship between independent and dependent variables by fitting a best line. This best fit line is known as regression line and represented by a linear equation Y= a \*X + b

In this equation:

Y – Dependent Variable

a – Slope

X – Independent variable

b – Intercept

**regularization method:**

The regularization procedure aims at avoiding the model to overfit the data and thus deals with high variance issues

we have two regularization method which are

lasso: The Lasso is a linear model that estimates sparse coefficients with l1 regularization.

Ridge: Ridge regression addresses some of the problems of Ordinary Least Squares by imposing a penalty on the size of the coefficients with l2 regularization.

We also have elastic net which is a linear regression model trained with both l1 and l2 -norm regularization of the coefficients.

## **Decision Tree:**

This is one of my favourite algorithm and I use it quite frequently. It is a type of supervised learning algorithm that is mostly used for classification problems. Surprisingly, it works for both categorical and continuous dependent variables. In this algorithm, we split the population into two or more homogeneous sets. This is done based on most significant attributes/ independent variables to make as distinct groups as possible.

## **kNN (k- Nearest Neighbors):**

It can be used for both classification and regression problems. However, it is more widely used in classification problems in the industry. K nearest neighbours is a simple algorithm that stores all available cases and classifies new cases by a majority vote of its k neighbours. The case being assigned to the class is most common amongst its K nearest neighbours measured by a distance function.

These distance functions can be Euclidean, Manhattan, Minkowski and Hamming distance. First three functions are used for continuous function and fourth one (Hamming) for categorical variables. If K = 1, then the case is simply assigned to the class of its nearest neighbour. At times, choosing K turns out to be a challenge while performing kNN modelling.

## **Random Forest:**

Random Forest is a trademark term for an ensemble of decision trees. In Random Forest, we’ve collection of decision trees (so known as “Forest”). To classify a new object based on attributes, each tree gives a classification and we say the tree “votes” for that class. The forest chooses the classification having the most votes (over all the trees in the forest).

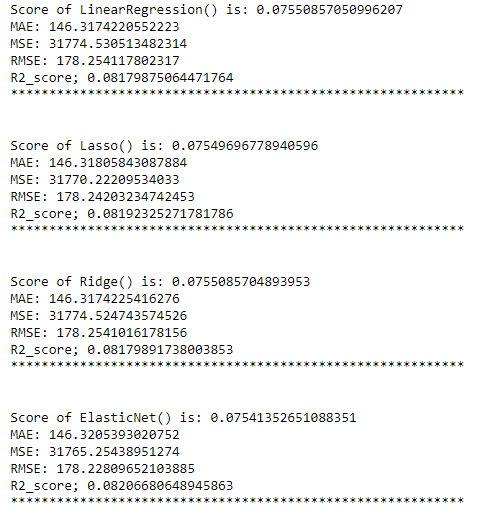
## **Gradient Boosting:**

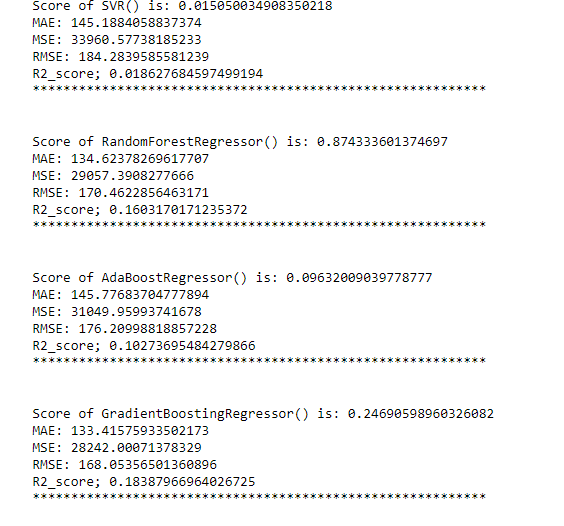
GBM is a boosting algorithm used when we deal with plenty of data to make a prediction with high prediction power. Boosting is actually an ensemble of learning algorithms which combines the prediction of several base estimators in order to improve robustness over a single estimator. It combines multiple week or average predictors to a build strong predictor.

Alright, so let’s go through this and see which of these models will perform the best and gives the best outcomes.



I am using For loop for all the models where I will print the accuracy score, mean absolute error, root mean squared error and r2 score of the data. Let’s see what is the outcome

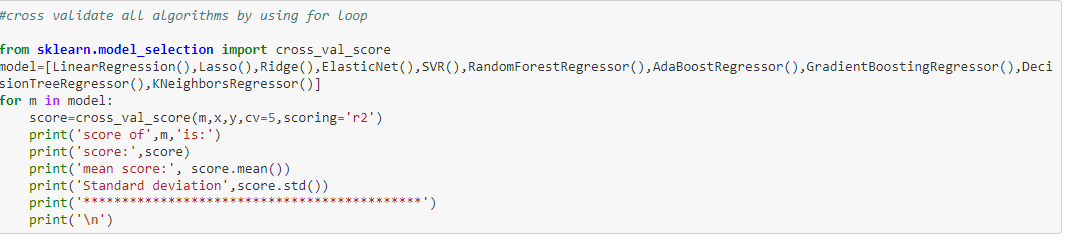




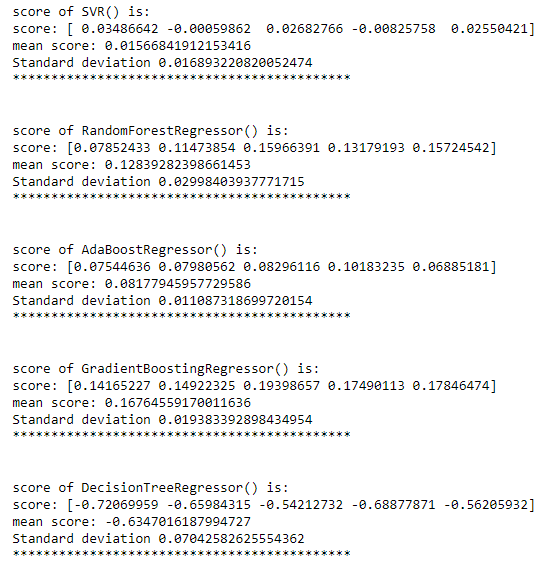
Now that we have loaded all the models, we can see which one is performing the best and which model is giving the best result. Since random forest regressor is working best and giving us the best result with the accuracy score of 87% we will use this model.

**Cross validation:**

 Cross validation is a method that is used to select a model that does not rely too much on the initial training set. I have used cross validation to see if there is any overfitting or underfitting of the trained dataset.

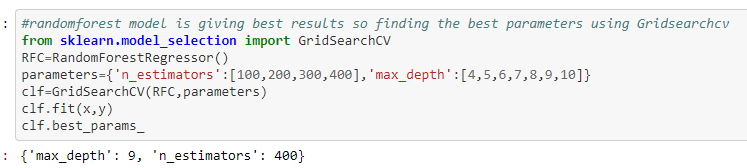


I have used For loop for cross validation as well and after cross validating the best result was given by random forest regressor.

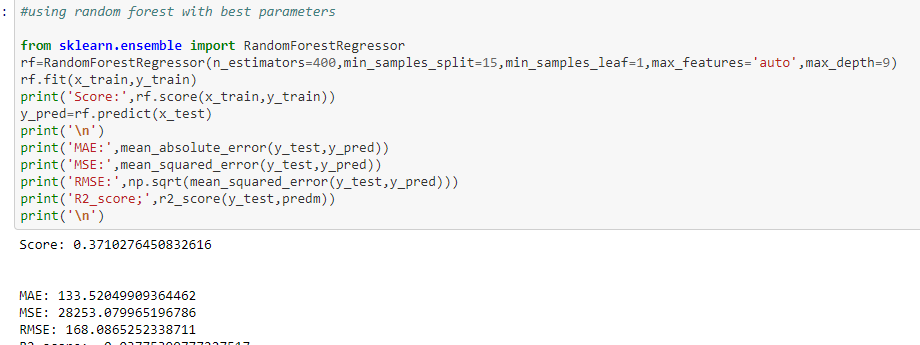


**Hyperparameter tuning:**

So, I am using hyperparameter tuning to find the best possible parameters to enhance the performance of our best performing model. I will be using grid search method as Grid search is arguably the most basic hyperparameter tuning method. With this technique, we simply build a model for each possible combination of all of the hyperparameter values provided, evaluating each model, and selecting the architecture which produces the best results

****

We have used GridSearchCV to find the best parameter for our best performing model -random forest regressor. I will be using this parameter to get the best possible outcome.



So, after all the process we are finally getting 37 % accuracy which is low but I will be using this because out of all the models it is the one which is performing the best and giving us the best results.

**Concluding remarks :**

I had tried to improve the accuracy score by changing the approach many times but this is the best possible outcome I got from this dataset. As I mentioned the final accuracy is not very good but I am ok with it because all the overfitting and underfitting is removed. I would like to summarize the whole blogs in my concluding remarks

1)there were two sets of data for this project I have used the trained dataset for the blog.

2)there are 7 columns in the dataset out of which 6 are object type and 1 is integer type

3)higher the qualification higher will be the doctor’s consultation fees.

4)ENT specialist and dermatologist are charging the highest fees.

5)there are null values present

6)In the model building random forest regressor is giving us the best result.

I have tried to make the blog as simple as possible so that anybody could understand it. I think this project serves well as an introduction to the basic concepts of Machine Learning and implementing training models. That’s pretty cool.

Blog by

Akchay pradhan.