**Introduction:**

We all have been in the situation where we go to doctor in our emergency or in our sickness, to know which doctor is best where we need to god where we can get a best treatment to all these we use the Google review get advice from others also we check the consultation fee and bed charges etc. for that we will ask the person who already gone there to take treatment but in every case in does not exist but if we have a data of the doctor and we can easily able predict the doctors consulting fee. Ad this can be done by the data scientist with building the best model with the data we have

**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

Miscellaneous Info: Extra information about the doctor

Fees: Fees charged by the doctor (Target Variable)

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

**Exploratory data analysis:**

The given dataset has 1987 rows and 6 columns and the given data consists of String and Float value so the dataset need to be changed either Float of Integer

The dataset have 4392 Null value in Rating and 3454 Null value in Miscellaneous Info and this has been checked using datasetname.isnull().sum()

To get some of the visualizations about the given data set I have used Univariate , bivariate and multivariate Analysis where we can get some ideas over the data actually what does the data holds and it will be quiet interesting using the plots like cat plot, Dist plot, Box plot, Hist plot, Heat map Boxen plot, Count plot etc.

The given data set contain two sources which is training datasets and Testing data set and both the data set have been combined to do a data preprocessing and EDA process

**Data Encoding**

The given data which is all consists of String value ad into should be converted into Float or string value for rating % symbol has removed by using replace command

The Miscellaneous Info consists of two parameters which is one with Feedback and location with Nan value in this the data with value has been replaced as 1 and Nan value has replaced with 0

The Place consists of city and Location and which has been separated has two columns and the Nan value has been replaced with Unknown because the location has mentioned

The experience columns consists of their major studies and In which they are specialist in And In which they has been separated and concerted in to Float value for further use

The same process has been followed in converting Experience data’s into Float value what we used to change the qualifications ,The year of experience has been removed from the data set using lambda function

Categorizing the dataset with fees which lies which lies in certain range and creating a another column to get the better predictive model

After treating all the elements the new data sets which consists of 5961 rows and 10 columns

To get better predictive model the skewness in experience has been removed from the Experience dataset, the experience alone has been treated because that has large informal Value in that

**Treating the Null Value**

While preprocessing the data all null value has been removed from Miscellaneous Info and still rating need to be treated to remove null value the Nan are replaced with -999

**Separating Test and train Dataset**

To make a model the test and train date set to be separated also to encode the data has float value

The train and test data set has been separated by using the columns we created Source

Once both the data set has been has been separated the additional columns like source in train datset and source and fees columns in test data set has been removed

**Encoding process**

To encode the data from String to float or string used the ordinal Encoder once concerting the data into float we experienced each columns had more differ from the each columns in values for that the MinMax Scalar has been used to fit the data and it has been only to the source variables

**Building Machine Learning Models.**

Since it is a regression model the algorithms like Linear Regression, RandomForestRegressor, DecisionTreeRegressor, KNeighborsRegressor has been imported When we are working with the regression type method the value or score should be focused on R2\_score, mean squared error, mean absolute error also to split the data the train\_test\_split dictionary has been used

The dependent variable and the target variable has been separated from the train dataset

To get the best predictive score the Best random need to ne found out for the given data set for that use the code like

maxAccu=0

maxRS=0

**for** i **in** range(1,100):

x\_train,x\_test,y\_train,y\_test=train\_test\_split(x,y,test\_size=0.20,random\_state=i)

rfr.fit(x\_train,y\_train)

pred=rfr.predict(x\_test)

acc=r2\_score(y\_test,pred)

**if** acc>maxAccu:

MaxAccu=acc

MaxRS=i

print("Best accuracy is",maxAccu,"On Random State", maxRS)

When using the different Algorithms got very good R2\_ score for all the data set so the Hyper parameter has been done to get more better R2\_score and get Better Cross validation score

On doing hyper parameter tuning found the Random forest Regressor performing well comparing to all other algorithms, and finally the best model has been saved as a pickle file for future use

