**Predict A Doctor's Consultation Fee**

AIM(Predict doctor consultant fee)

We have all been in situation where we go to a doctor in emergency and find that the consultation fees are too high. As a data scientist we all should do better. What if you have data that records important details about a doctor and you get to build a model to predict the doctor’s consulting fee. So that dataset are use to predict the consultation fee of doctor and also indicating feature reflected that so trying to find out the answer and the predicted result of that dataset.

Feature

* 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.

Feature creation

1. Qualification columns has all the study records of a doctor in a string format by comma separated. E.g. MBBS, MS - Otorhinolaryngology. So, it makes sense to create features out of it like Diploma, Bachelor, Masters, and Extra study, etc., but the problem is we have more than 600 unique study and major combination and we need understanding of all these Medical course to created different features, that's why I have TfIdf vector and TurncatedSVD on these features to create final 20 features out of this Qualification information.
2. Extracted just number out of Experience.
3. Removed % from Rating.
4. More than 3000 records doesn't have Miscellaeous\_Info and in majority case it only consists info about Experience and location, which we already have, so I am not using this column in model. I have created an extra column Has\_M\_Info which indicates if the record has miscellaeous info or not.
5. Place column has Area and City info comma separated, so I have created two columns Area & City out of it.

EDA

* After doing data analysis it is clear that Rating and Experience has less correlation with target variable Fees.
* It was clear from Box plots of Fees for Profile and City that these columns are more important for segregation.
* But without adding Education information we can still not find good separation in data groups.

Evaluation Metrics

* Mean squared error (MSE) is used as a loss function and RMSLE as a evalution metric.
* Profile and City is used as categorical variables.
* Used  **Cross validation** and averaged predictions from each fold for test dataset for final submission.

Improvement and achieve result

* Achieved Validation RMSLE of 0.625 and RMSE of 168.87
* If we created more features out of Experience column by understanding of all medical courses, then there is a high chance to improve our model performance.