**Project-MovieLens Case Study .**

After getting brief understanding about python and data science concepts based on the flow from business objective to deployment. First tried to go through what is the data all about how many features does it have .

**First Step-1** Business Problem and Data Acquisition.

The case study has three datasets using them we have to perform analysis the datasets are users.dat movie.dat and ratings.dat also find which model has got good accuracy and find which features affect ratings.

**Step 2-** Data preprocessing and wrangling.

Data is imported jupyter notebook after importing merge the data into new data then checked the size and shape of data.

Shape: 8 columns x 1000209 rows

Features: UserID, MovieID,Movie-Name,Genre,Ratings,Timestamp, Pincode,Occupation.

UserID: ranges from 1 to 3500

MovieID: ranges from to 3495

Ratings: ranges from 1 to 5

Occupation: ranges from 1 to 19

Age:ranges from 1 to 56

On applying wrangling techniques the data doesn’t have any null values.

**Step 3:** Exploration.

Explored data using pandas, numpy, matplotlib libraries.

**Step 4:** Model development.

Data has been divided into train test with ratio 50:50 i.e. 50% training 50% testing.

Using logistic regression: accuracy is 35%

Using Decision Tree: accuracy is 100%

Using random forest regression: accuracy is 99%

Comparing all three algorithms decision tree suffers from the problem of overfitting hence random forest gives better result.

For rest kindly refer source code.