**D.E.Code**

**Helping an investment bank**

**Assumptions & Modelling Choices:**

* Since it is a regression problem in which the test data contains merely 2 features namely, Advisor\_Id and Investment\_Id, the scope for using linear regression rules out and main emphasis is on these 2 features only. In such a situation, the most suitable algorithm turns out to be the K-Nearest Neighbour Regression (KNeighborsRegressor) algorithm. K-NN is a memory-based approach in which the classifier immediately adapts as we collect new training data.
* Initially, I modified the training data by changing the ‘Date’ column to include just the Month because the year is same for all training examples.
* I also added a new column named ‘Date’ to the given test data and filled it with 7 because the test data is for July (7).
* From the analysis of data it comes out that some of the Advisor\_Ids are negative which makes no sense, so I have taken only the numerical (absolute) value of all Advisor\_Ids.
* Initially, I tried to train my model with only 3 features: Advisor\_Id, Investment\_Id and Date (in fact month) neglecting the remaining features because no such significant correlation can be established among them by drawing the Pearson Correlation heatmap. I took algo=’auto’, weights=’distance’ and n\_neighbor=HIGH\_VALUE for this model because:
  + algo=’auto’ chooses the best fitting algorithm out of ball tree and kd tree.
  + weights=’distance’ gives higher preference to the closer points.
  + Features are less so I have used more neighbours for achieving accuracy. Moreover, the boundary becomes smoother with increasing value of K. Large K also means noise reduction and high bias model.
* I tried to generate an output and test it for AUC-ROC accuracy. It was around 0.66.
* On further analysis, I decided to drop Date (Month) feature for training my model for prediction and by doing so accuracy increased.
* Moreover, I also tried to normalise data because the Advisor\_Id and Investment\_Id are too large but it also reduced the accuracy of my model. So I decided to drop this normalisation also.

**Tools & Libraries:**

* Sklearn library of python
* Pandas library of python

**NOTE:** I am a beginner in machine learning so I lack in in-depth knowledge of feature engineering and data analysis.