**Approach for the Problem Statement**

(Link to the problem - <https://www.kaggle.com/c/nyc-taxi-trip-duration> )

* First of all after getting the data I analyzed the format, datatypes of all the variables, identified the dependent variable.
* Then basic pre –processing like checking for the missing value, extracting week and month information from month variable, creating extra variables like distance from the two latitude and longitude points, speed and many more.
* Basic exploratory data analysis and visualizations of independent variables and relation of independent variables with dependent variable i.e. trip\_duration to gain some insights and make some quick decisions.
* During this analysis various outliers or noise in some row were identified like data points having distance greater than 500 km, trip\_duration greater than 20 hours and many more. These all data points were removed from the training set.
* In order to implement Xg-Boost model factor variables like month, weekdays were hot encoded to many numeric binary variables.
* Merged the existing training data with external weather data of New York City in order to include the effect of weather.
* Final training set was developed taking all the necessary variables.
* Finally checked for the co-relations of the variables and later removed extra co-related independent variables in order to create more robust model.
* Data wrangling for test data in similar format.
* Now, finally applied Xg-Boost model in which hyper-parameters were tuned by the random search over the range of the values of different parameters specified 100 times and then cross-validating each search 5 times.
* “Rmse” was used for measuring training error and “mse” for average validation error.
* At last Feature importance was done for deciding the most important features for this model.
* Finally the model predicted trip duration for test data.

**Gaining accuracy with more time over Super computers!!**

* Let say If I had more time I surely had invested more time on features research and creating more variables effecting time duration like marking sites in New York city on traffic jam conditions and checking whether sites are present in the travel route, including road type, condition variable, checking for any events in city on the day of travel and searching out more variables like this which could effect.
* If we have more strong machines or let say a Super computers we can go with many iterations in order to find most fit parameters.
* We can also implement vast grid search algorithm instead of random search for tuning parameters which otherwise is very time consuming.
* May be checking for all other methods like training neural network model and lastly choosing the best.