**GiveMeSomeCredit**

**Question:** Given a set of financial features, predict what are the chances that someone will experience financial distress in the next 2 years.

**Data Acquisition:** Data obtained from <https://www.kaggle.com/c/GiveMeSomeCredit/data>.

**Data Cleaning:**

There were NA values in 'MonthlyIncome' and 'NumberOfDependents'. These may be missing data. I decided to replace NA values of 'MonthlyIncome' with it's median and NA values of 'NumberOfDependents' with 0.

'NumberOfTimes90DaysLate' and 'NumberOfTime30-59DaysPastDueNotWorse' had values 96 and 98. If we think about it, this seems to impossible since 96\*90/365 ~ 24 years. When I further checked such records, I saw that 'RevolvingUtilizationOfUnsecuredLines' was 0.999 for all such records. These must be typographical errors in data. Since the count of all such records was very less as compared to our whole sample count, I decided to drop such records from our analysis.

**DATA EXPLORATION, MODEL FITTING AND PREDICTION**

Simple Data exploration was generated using info() and describe() of dataframe.

**Random Forest Regression** was applied on the data using sklearn’s RandomForestRegressor.

Basic tuning of the model was done while playing with n\_estimators and min\_samples\_leaf.

We got AUC- ROC score as 0.834119567515 for the following parameters:

{'warm\_start': False, 'oob\_score': True, 'n\_jobs': 1, 'verbose': 0, 'max\_leaf\_nodes': None, 'bootstrap': True, 'min\_samples\_leaf': 1, 'n\_estimators': 200, 'min\_samples\_split': 2, 'min\_weight\_fraction\_leaf': 0.0, 'criterion': 'mse', 'random\_state': None, 'min\_impurity\_split': 1e-07, 'max\_features': 'auto', 'max\_depth': None}

Using this tuned model, we went on and predicted probability for data of cs-test.csv and generated **submission.csv** as per the format of sampleEntry.csv.