RealData.csv

The data provided was skewed as 80% of cases had BAD=0 and 20% of cases had BAD=1.
For the above reason , i had replicated the 20%(BAD=1) data 3 times to remove the skewness of the data - using EXCEL.
Few columns having categorical values were split into several columns of binary values. I had proceeded to do the same using label_binarizer , but faced difficulty in proceeding and did the same using EXCEL.
The data attached contains the above modifications done on it.

variableSelector.csv

The data was imputed using MICE technique. Multivariate imputation by chained equations Creating multiple imputations, as opposed to single imputations, accounts for the statistical uncertainty in the imputations.
I had selected variables using RFE - The Recursive Feature Elimination (or RFE) works by recursively removing attributes and building a model on those attributes that remain. The variables were ranked and required variables were selected.
['LOAN', 'MORTDUE','REASON' , 'VALUE','DELINQ', 'DEROG' ,'CLAGE','Other','DELINQ', 'Office' ,'Sales', 'ProfExe'] was the columns selected to build the prediction tool on.

Tool.py

Decision tree was implemented with maximum depth = 12 , minimum samples in leaf = 200 and minimum sample split = 500. The accuracy was tested on test data . This tree was converted to code so as to build a prediction tool.

Classification_tool.py

The decision tree was coded hence building a tool to predict BAD for a new input of required variables.

Tree.png

Decision Tree build based on the data