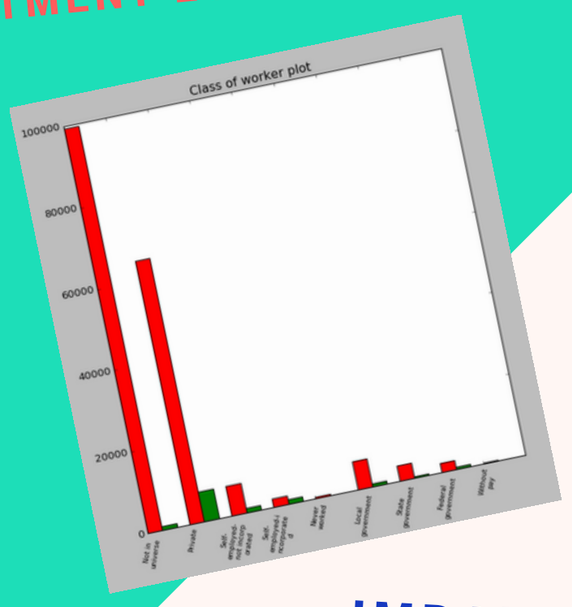
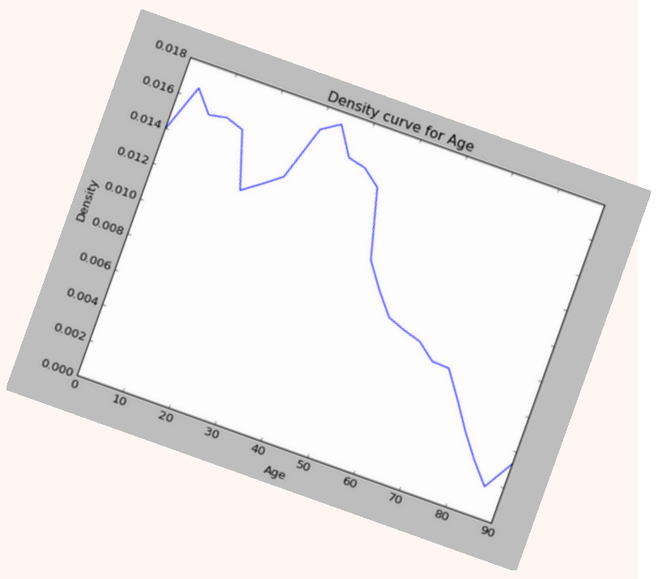
PREDICTIVE MODEL FOR CENSUS DATA

A PREDICTIVE MODEL TO DETERMINE THE INCOME LEVEL FOR PEOPLE IN US. WITH INCOME LEVELS BINNED AT BELOW 50K AND ABOVE 50K.
BASED ON AGE, MARITAL STATUS, INCOME, FAMILY MEMBERS, NO. OF DEPENDENTS, TAX PAID, INVESTMENT ETC





IMBALANCED DATA, WITH MAJORITY CLASS PROPORTION OF 94%
SAMPLING TECHNIQUES APPLIED
DRAWBACKS :

- UNDERSAMPLING LEADS TO LOSS OF INFORMATION
- OVERSAMPLING LEADS TO OVERESTIMATION OF MINORITY CLASS

ASSIGNING WEIGHT TELLS THE ALGORITHM THAT THIS (MINOR-ITY) CLASS IS MORE IMPORTANT. THE OTHER CLASS MAY BE PREDICTED AS WELL AS POSSIBLE, BUT THE MINORITY CLASS HAS TO BE PREDICTED WITH FULL CERTAINTY SINCE IT IS GIVEN HIGHER WEIGHT.

CONCLUSION:

GENERAL EXPECTED PERFORMANCE IS 94% ANALYSED USING NAIVE BAYES, LOGISTIC REGRESSION, DECISION TREES, RANDOM FORESTS, SVM, XGBOOST.
BEST RESULT ACHIEVED THROUGH XGBOOST!

XGB Classifier	precision	recall	f1-score	support
Less than 50k	0.96	0.99	0.98	93576
More than 50k	0.75	0.35	0.47	6186
avg / total	0.95	0.95	0.94	99762

XGB Classifier Smote 0.4	precision	recall	f1-score	support
Less than 50k	0.97	0.97	0.97	93576
More than 50k	0.56	0.54	0.55	6186
avg / total	0.94	0.95	0.94	99762

ABOUT XGBOOST

High speed
Regularization
Parallel Processing

High Flexibility

Handling Missing Values

RAJU KHANAL (110849511)
DEVESH SISODIA (110951296)
AKASHA ROY (111121421)