Predicting earning potential on Adult Dataset

Objective

The objective is to create a classification model which can **predict individuals whose salary exceeds fifty thousand US dollars** by modeling census data containing demographic information such as age, gender, and education level and employment type.

Data Set and Metrics to be used:

The US Adult Census dataset is a repository of 48,842 entries extracted from the 1994 US Census database. Target income is imbalanced with income less than 50k being 75.2% and income greater than 50k being 24.8%. As we are focusing on the income greater than 50K we should take it into account which metric to look for while modelling. **I considered accuracy, ROC AUC and Matthew Correlation Coefficient as metrics to be observed to tell whether the model is good.**

Steps: Data is cleaned, explored, imputed for missing values and transformed such a way that it gives good result while modeling in less time. From exploratory analysis we found that.

- Income has positive correlation with age, education-num, sex, capital gain, capital loss and hours per week where as negative correlation to marital status and relation ship. Also relationship and hours per week is negatively correlated.....so on.....
- More education does not result in the same gains in income for Asian Americans/Pacific, Black, American Indian Eskimo but it increase in white and other
- More education does not result in the same gains in income for women compared to men.
- Older people make more, except for Asian Americans/Pacific Islanders.

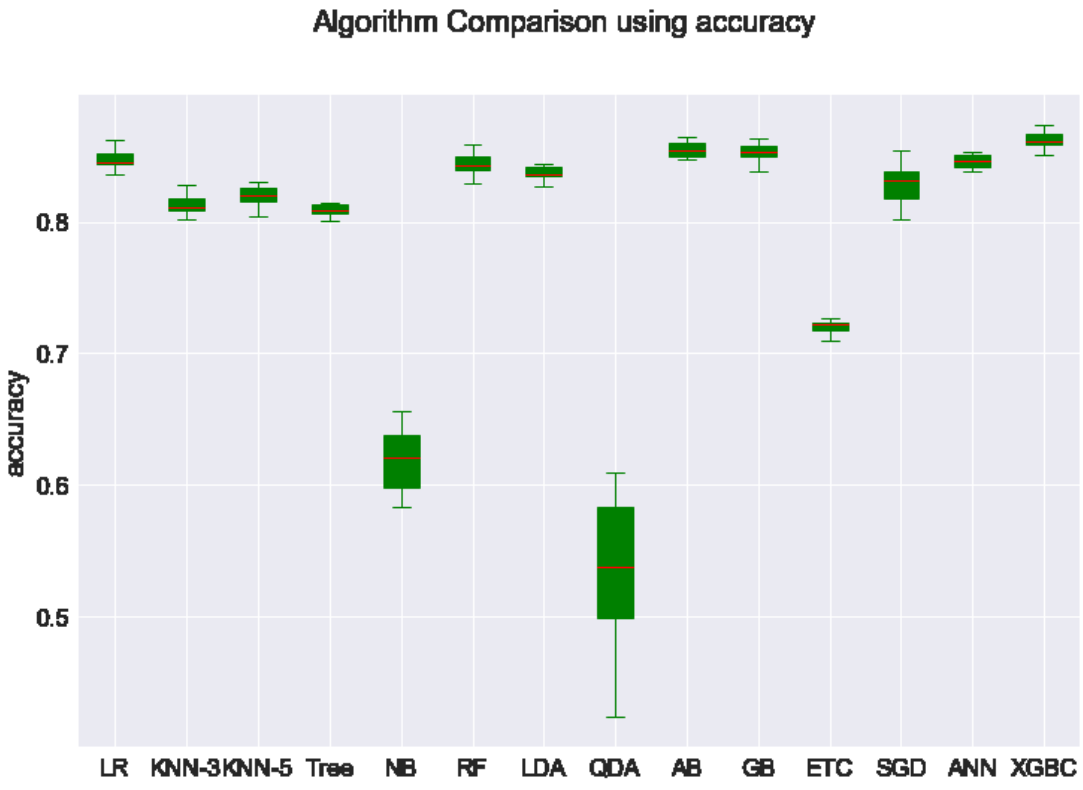
Modeling: For modeling I considered various algos and found the metrics accuracy Roc Auc and Matthew Corr. Coeff. We considered other metric other than accuracy as we have imbalanced target and accuracy won't be good metric. It is better to check Roc Auc and Matthew Corr. Coeff (explanation is on notebook why they are better metric).

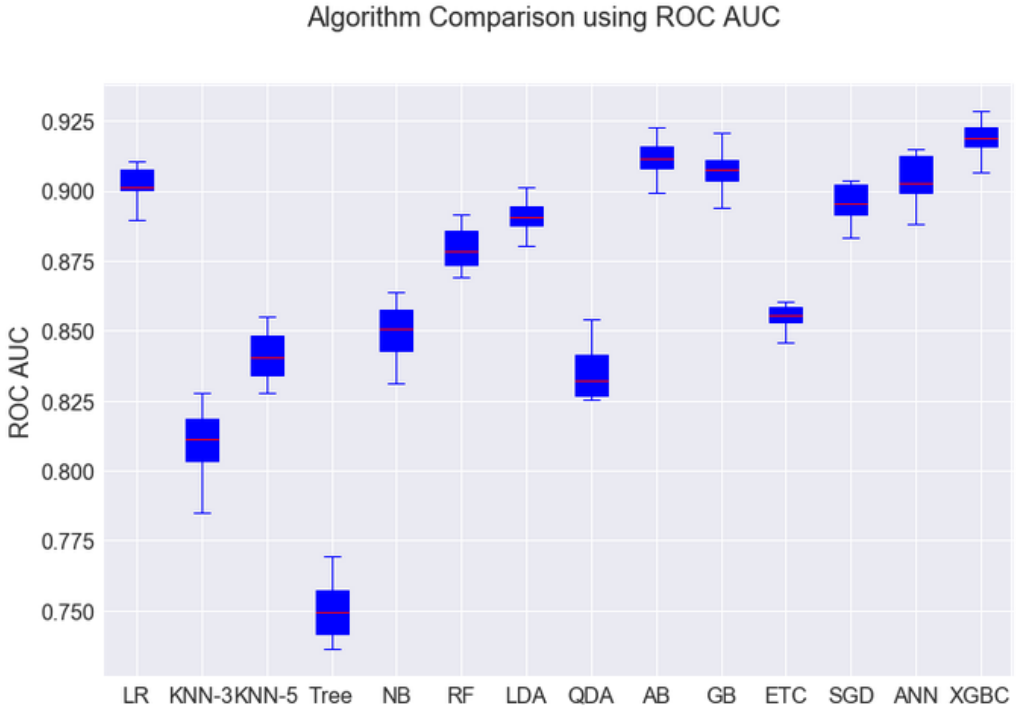
	model	accuracy	ROC-AUC	Matthew Corr. Coeff	
0	LR	0.847407	0.901746	0.564248	'LR', LogisticRegression
1	KNN-3	0.811564	0.809315	0.492722	KNN-3', KNeighborsClassifier
2	KNN-5	0.820470	0.840926	0.511583	KNN-5', KNeighborsClassifier
3	Tree	0.811470	0.751074	0.495447	Tree', DecisionTreeClassifier
4	NB	0.618099	0.849665	0.399789	NB', GaussianNB
5	RF	0.843714	0.879477	0.557762	RF', RandomForestClassifier
6	LDA	0.837787	0.890294	0.543441	LDA', latent Dirichlet allocation
7	QDA	0.534275	0.830616	0.335134	QDA', *Quadratic discriminant analysis*
8	AB	0.855073	0.911220	0.590134	AB',AdaBoostClassifier(DecisionTreeClassifier
9	GB	0.852466	0.907344	0.598388	GB',GradientBoostingClassifier
10	ETC	0.720262	0.854511	0.461091	ETC',ExtraTreesClassifier
11	SGD	0.829968	0.895163	0.489755	SGD',SGDClassifier
12	ANN	0.847221	0.904053	0.577683	ANN',MLPClassifier
13	XGBC	0.862211	0.918199	0.597795	'XGBC', XGBClassifier

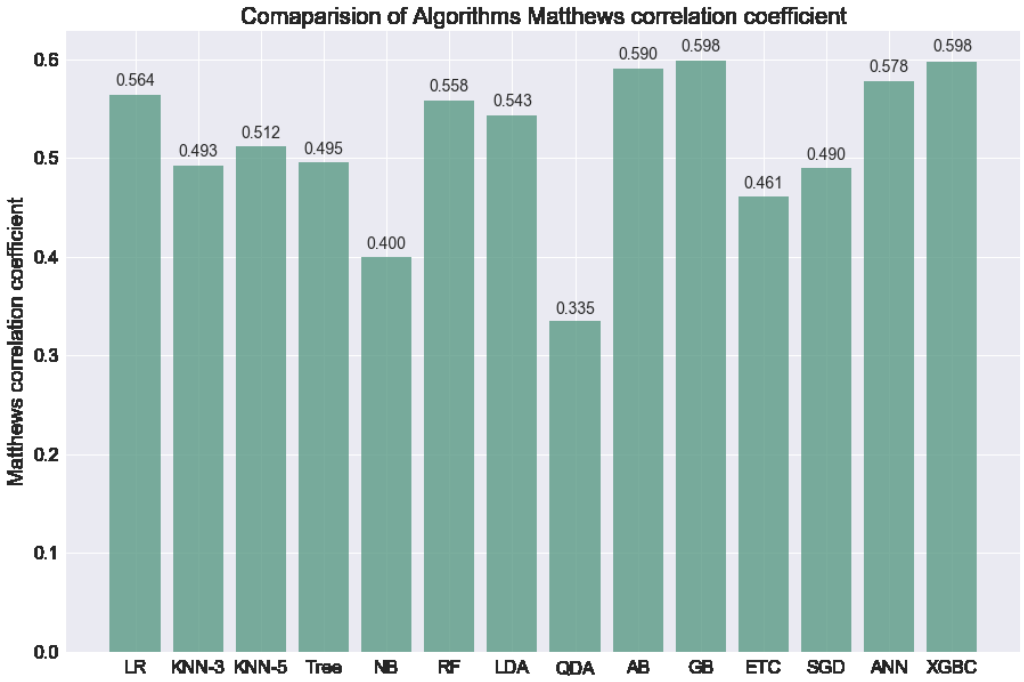
For Accuracy and Roc_auc the modeling was cross validated so as to avoid outliers. Support Vector machine (linear) accuracy output was 82% which was carried separately .Three best models got(AB, GB & XGBC) from the above is taken and parameter tuning considering the metric as ROC_AUC was done (all together 25 parameter tuning). I was lucky that I had good parameters initially so I got result for ROC_AUC same as above. Then further I tried to stack the AB+GB+XGB and see ROC_AUC improves or not. It was ROC_AUC to be 92% same as that of XGBC. Hence I selected the XGBC to represent the model.

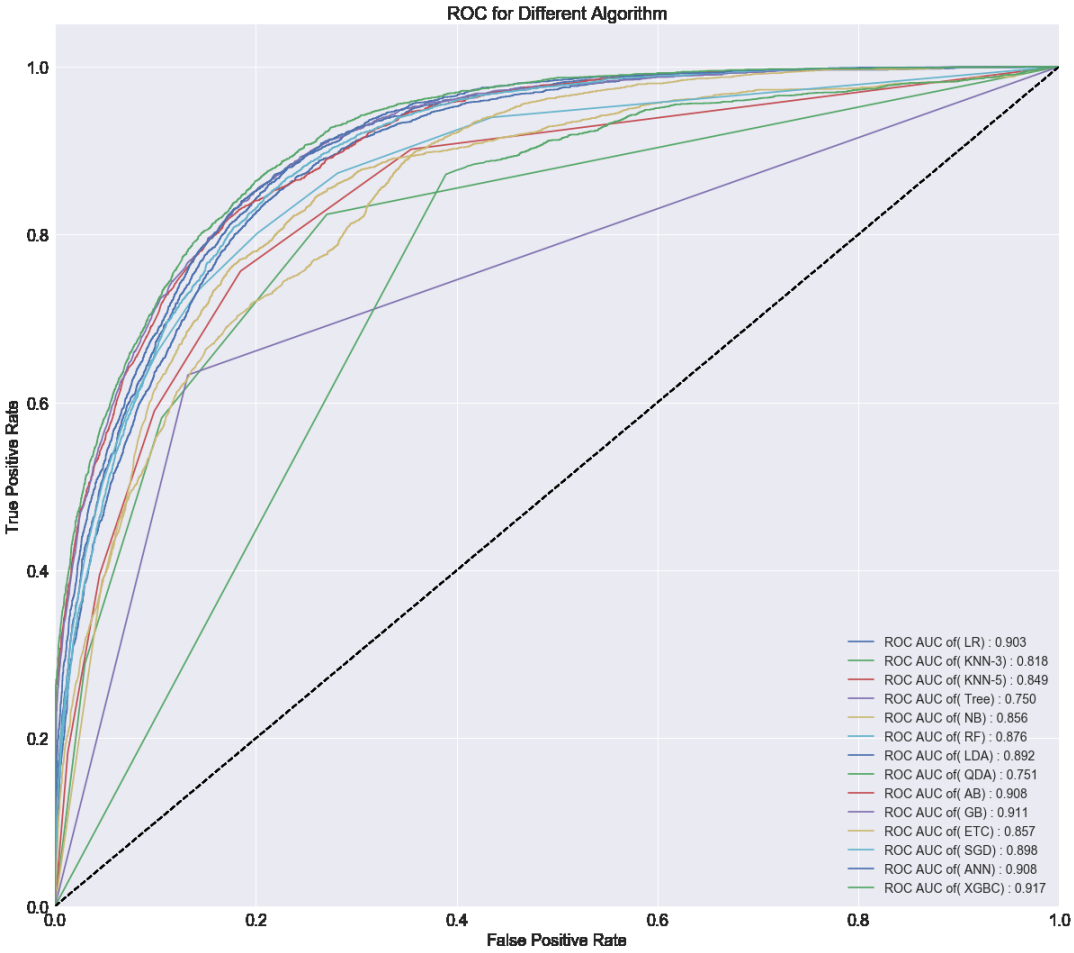
Results: Out of 50 modeling done, Xgboost Classifier is the winner of all models with accuracy 86%, ROC_AUC=92% and Matthew Coerr.Coeff=0.6.

Appendix


Algorithm Comparison using accuracy
accuracy
0.8
0.7
0.6
0.5
LR
KNN-3
KNN-5
Tree
NB
RF
LDA
QDA
AB
GB
ETC
SGD
ANN
XGBC


Algorithm Comparison using ROC AUC
ROC AUC
0.925
0.900
0.875
0.850
0.825
0.800
0.775
0.750
LR
KNN-3
KNN-5
Tree
NB
RF
LDA
QDA
AB
GB
ETC
SGD
ANN
XGBC


Comaparision of Algorithms Matthews correlation coefficient
Matthews correlation coefficient
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.564
0.493
0.512
0.495
0.400
0.558
0.543
0.335
0.590
0.598
0.461
0.490
0.578
0.598
LR
KNN-3
KNN-5
Tree
NB
RF
LDA
QDA
AB
GB
ETC
SGD
ANN
XGBC


ROC for Different Algorithm
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC AUC of(LR) : 0.903
ROC AUC of(KNN-3) : 0.818
ROC AUC of(KNN-5) : 0.849
ROC AUC of(Tree) : 0.750
ROC AUC of(NB) : 0.856
ROC AUC of(RF) : 0.876
ROC AUC of(LDA) : 0.892
ROC AUC of(QDA) : 0.751
ROC AUC of(AB) : 0.908
ROC AUC of(GB) : 0.911
ROC AUC of(ETC) : 0.857
ROC AUC of(SGD) : 0.898
ROC AUC of(ANN) : 0.908
ROC AUC of(XGBC) : 0.917

Confusion Matrix of Various Classifiers