TECHGIG

CODE

GLADIATORS

[Loan Eligibility Process]
[BOT Masters]

Team Detail

Insight
Harshavardhan KP

Key
Harshavardhan KP

Drive
Harshavardhan KP

Testing
Harshavardhan KP

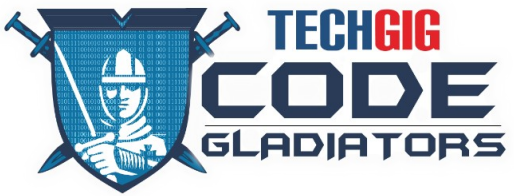
Develop
Harshavardhan KP

Performance Tune
Harshavardhan KP

Objective

A.Objective of this excercise is to analyse existng loan data and build a predictive model which will predict applicatns who could be eligible for Loans.

B. This prediction will be used by XYZ bank to target customers based on their eligibility



Approach

1. Identify dependent and independet variables
2. Perform the goodness of fit (Chi square , LDA , correlation) test to decide suitable variables that affect loan approval decision
3. Missing value treatment and outlier treatment
4. Data analysis (univariate and multivariate analysis) to check the distribution of data
5. split training data in to training sample and test sample
6. Build Linear and non linear models (CART, XGBoost, LDA, LR, NB,KNN,SVM)
7. Compare their performance through cross validation score
8. chose the top 3 models for prediction
9. predict the outcome of test data
10. take mode value of all 3 predictions as final prediction (This will reduce errors and increase accuracy of prediction)

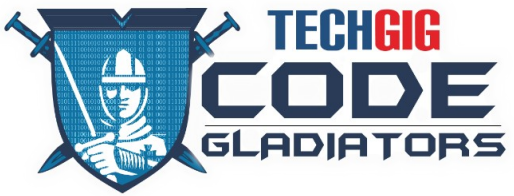
TECHGIG CODE GLADIATORS

Build Tools

Code is in python scripts it has to be executed in environment with following configuration

1.Python2.7 with following modules (sklearn,pandas,Xgboost,math,numpy)

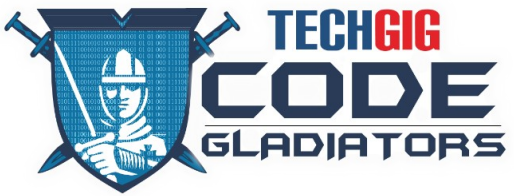
2.Ubuntu (14.4)



Source/Build/Execution Guidelines

Below steps should be followed to generate the result file

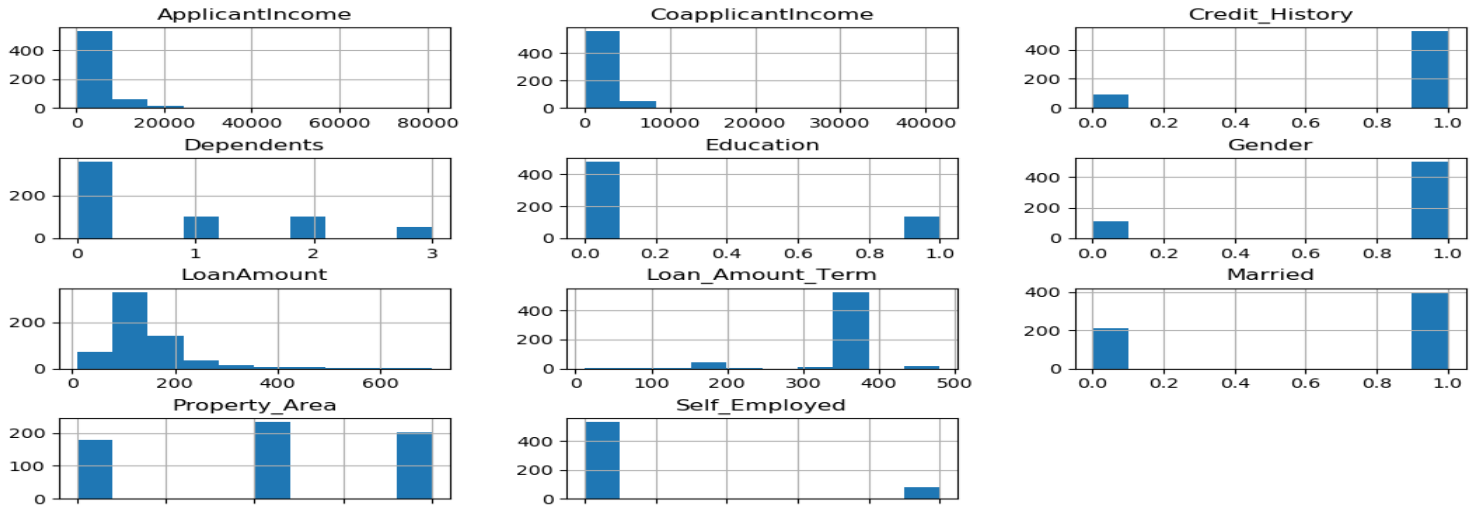
1.Downlod the script to a folder
2.Copy the data file to same folder
3. Open command prompt and navigate to above foder
4. Run the script file using following command “python LEP.py”
5. Result file “submission.csv” will be generated with results



Data Exploration and Data Cleaning

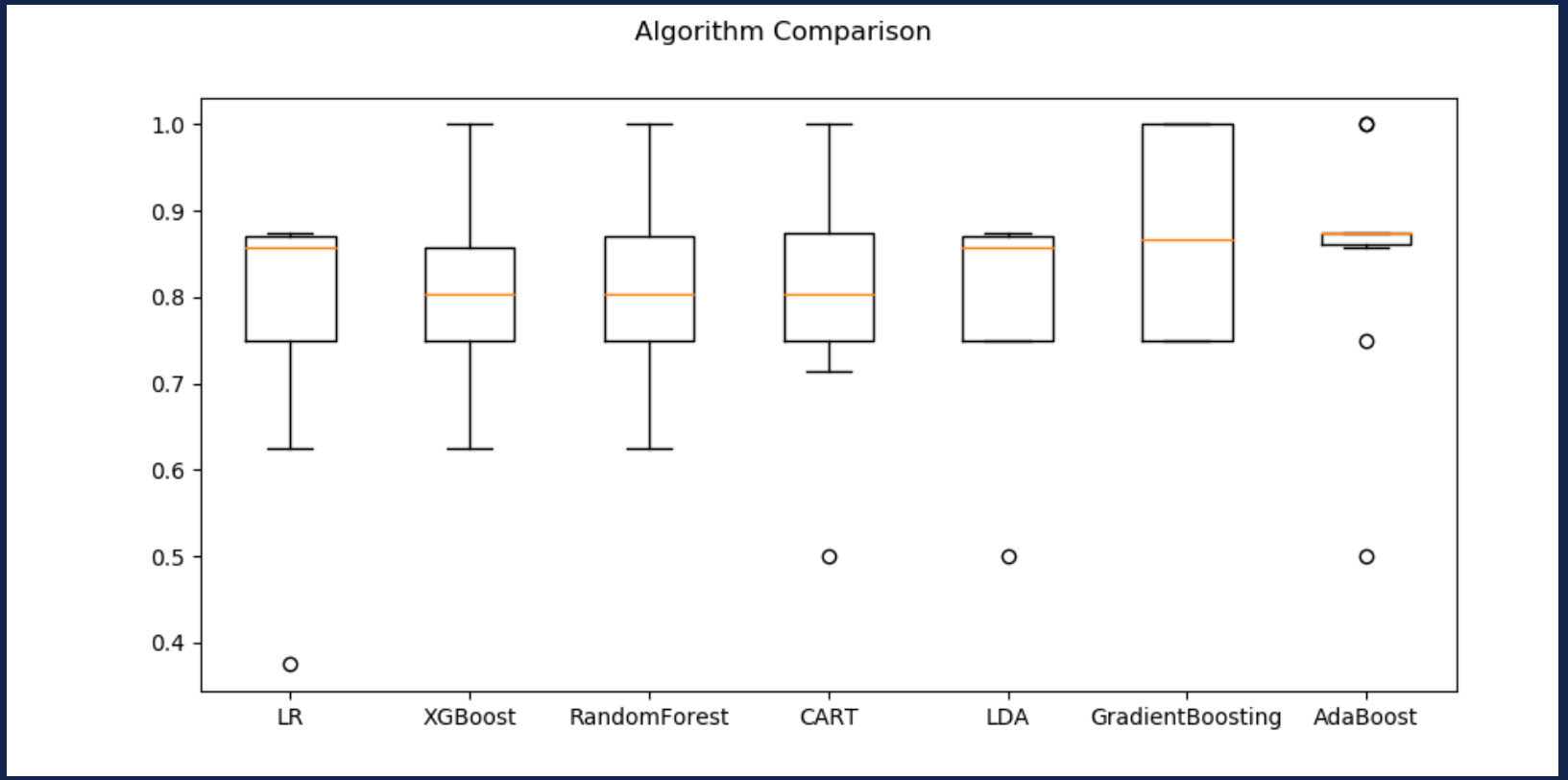
A. Null values and outliers are imputed with following rule:: Median value for numeric data and mode value for non numeric
B. New feature NetIncome is created ,which consolidates income of applicants and fecilitate in better predictions

Univariate plots :Histogram



TECHGIG CODE GLADIATORS

Modelling and Validation



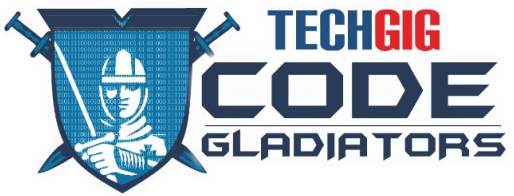
TECHGIG CODE GLADIATORS

Modelling and Validation

A. Collaction of Linear and non linear classification models were chosen for modelbuilding. (Logistic regression, LDA,Decision tree ,Xgboost,SVM ,GaussianNB,

B. After Crossvalidation test with 10 splits of input data we observed that Decision tree, LDA and XGBoost had more accurate and consistent predictions , hence those three were chosen for final model

C. Final model precitions = mode (CART Prediction, XGBoost Prediction, LDA Prediction.)



Score and Evaluate Model

	Predicted	
Actual	T	F
T	7	2
F	1	13

	precision	recall	f1-score	support
N	0.88	0.78	0.82	9
Y	0.87	0.93	0.9	14
avg / total	0.87	0.87	0.87	23

TECHGIG CODE GLADIATORS

Possible Improvement

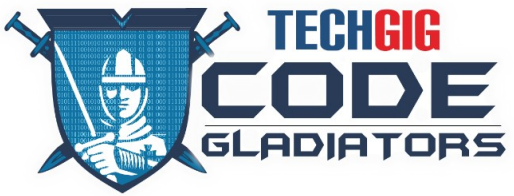
01 Decision Tree classification model can be furher tuned to get more accurate Predictions

02 Current modeluses random sampling to train . If stratified sampling is used more accurate predictions can be made for example : Single urban female, Single urban male, Single earning female in rural , Single earning male in semi urban

TECHGIG CODE GLADIATORS

Final Results / Summary

Final Predictions will be provided in a seperate file



Thank You

[Loan Eligibility Process]
[BOT Masters]