Project Report

on

Consumer categorization based on Family Income

Group No: 01

Shinakshi Sankhyan
Viraj Majalkar
Harshad Punghera

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF CYBER SECURITY
SCHOOL OF TELECOMMUNICATION

Problem Statement:

Reserve Bank of India collects various financial information from public and private sector banks as well as other govt entities. Based on information RBI publishes various financial and economic report for public information. One of such report is Consumer Survey Index which provides information on income levels of Indian consumers.

Based on consumer survey data available, considering multiple factors (features) like No of Earning Members in the Family vs No of Family Members, consumers Annual Income is to be categorized into different segment like "Below Poverty Line", "Low Income"," Mid Income" etc.

Overall summary of your solution

As Logistic Regression supports classification of data only in 2 clusters, approach of categorizing consumers into 3 different segments have been changed and now consumers Annual Income would be instead categorized into 2 different segments i.e. Below Average Income and Above Average Income.

Annual Income below Rs.3 lacs has been categorized as Below Average Income and above Rs.3 lacs has been categorized as Above Average Income.

Classification techniques used to solve the problem are mentioned below.

1. Naïve Bayes
2. Decision Tree
3. Logistic Regression
4. KNN
5. Support Vector Machine
6. Random Forest

Almost all classification techniques have resulted in similar confusion matrix and accuracy percentage.

Detailed description and analysis of solution

Data downloaded from RBI website using web scraping is imported in python to fir data analysis. Before any processing, data has total 13 columns and 5351 records.

During data analysis it has been observed that apparat from 2 variables Lower and Upper Range rest all variables contain null values. All such null values have been replaced with NaN.

As Annual Income is to be divided into 2 classes, income below 3 lacs have been replaced with value 0 (4710 records) and income above 3 lacs have been replaced with value 1 (641 records).

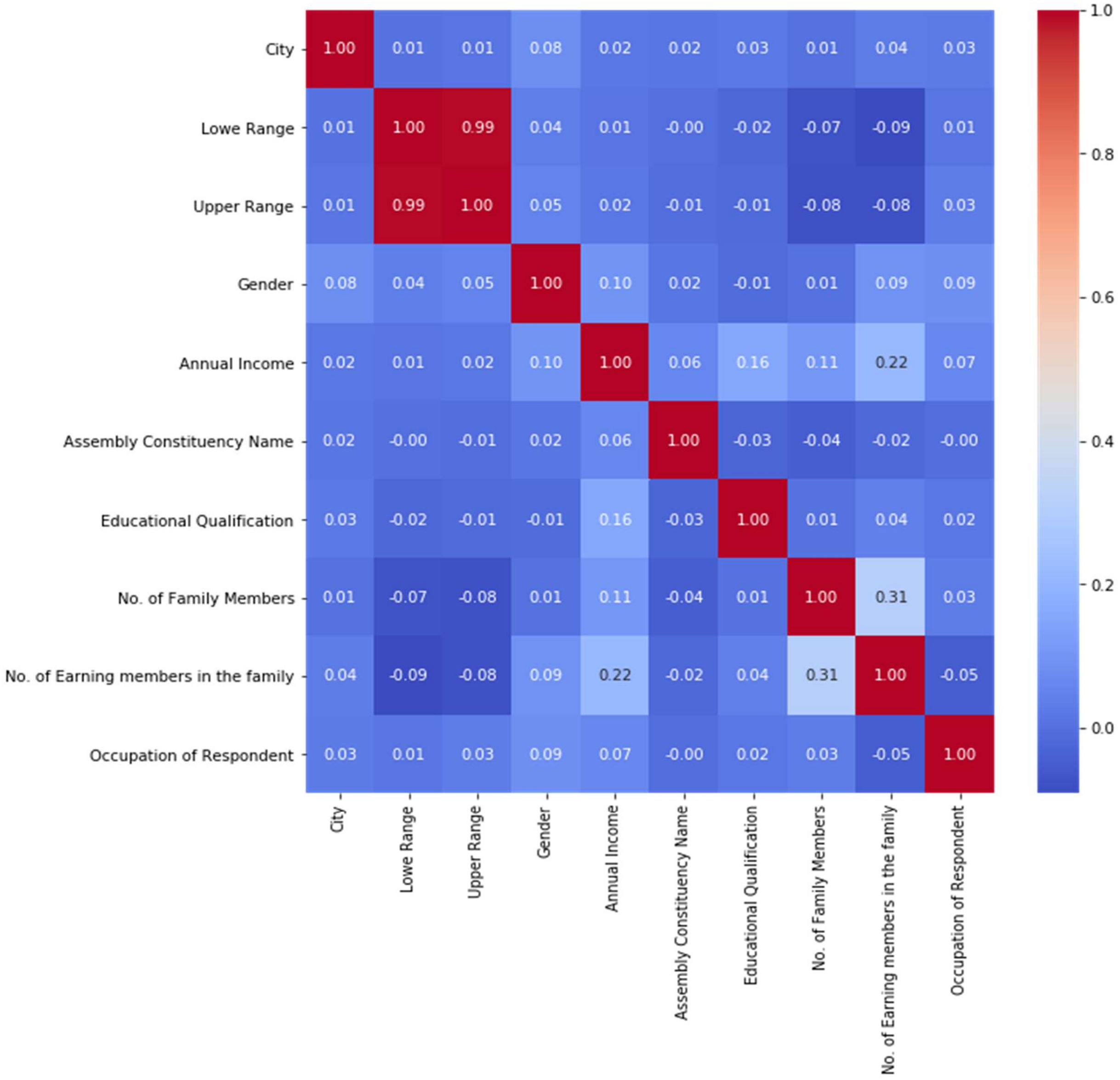
Data contains 7 columns having categorical values namely (Gender, Annual Income, City, Occupation of Respondent, No. of Family Members, Educational Qualification, Assembly Constituency Name, No. of Earning members in the family). Label encoding technique has been used to convert the categorical values to type numeric.

Correlation Matrix and *Heatmap Graph* has been used to understand correlation between the variables and it has been found that variable No of Earning Members in the Family is highly correlated to Annual Income.

Correlation Matrix

	City	Lowe Range	Upper Range	Gender	Annual Income	Assembly Constituency Name	Educational Qualification	No. of Family Members	No. of Earning members in the family	Occupation of Respondent
City	1	0.006975	0.010965	0.080782	0.02173	0.015449	0.027746	0.006253	0.036361	0.031308
Lowe Range	0.00698	1	0.994906	0.039919	0.01417	-0.00061	-0.019711	-0.073695	-0.091495	0.01498
Upper Range	0.01097	0.994906	1	0.051923	0.02057	-0.005111	-0.013393	-0.075572	-0.077507	0.033407
Gender	0.08078	0.039919	0.051923	1	0.09964	0.017443	-0.009126	0.006457	0.092385	0.085895
Annual Income	0.02173	0.01417	0.020567	0.099635	1	0.059973	0.157326	0.105107	0.218136	0.06601
Assembly Constituency Name	0.01545	-0.00061	-0.00511	0.017443	0.05997	1	-0.025868	-0.041066	-0.015252	-0.00467
Educational Qualification	0.02775	-0.01971	-0.01339	-0.00913	0.15733	-0.025868	1	0.010479	0.044843	0.022104
No. of Family Members	0.00625	-0.0737	-0.07557	0.006457	0.10511	-0.041066	0.010479	1	0.313214	0.034868
No. of Earning members in the family	0.03636	-0.0915	-0.07751	0.092385	0.21814	-0.015252	0.044843	0.313214	1	-0.047332
Occupation of Respondent	0.03131	0.01498	0.033407	0.085895	0.06601	-0.00467	0.022104	0.034868	-0.047332	1

Heatmap



Variables No of Earning members in the family, No of Family Members, Educational Qualification, Gender have been considered as predictor or independent variable and variable Annual Income has been considered as response or dependent variable. While applying all six techniques data has been split into train and test data set with 80:20 ratio.

Naïve Bayes

GaussianNB module of SKLEARN library is used to perform classification. Model has been trained approx. 4280 records and tested on 1070 records. With validating test records model predicts 86.74% accuracy.

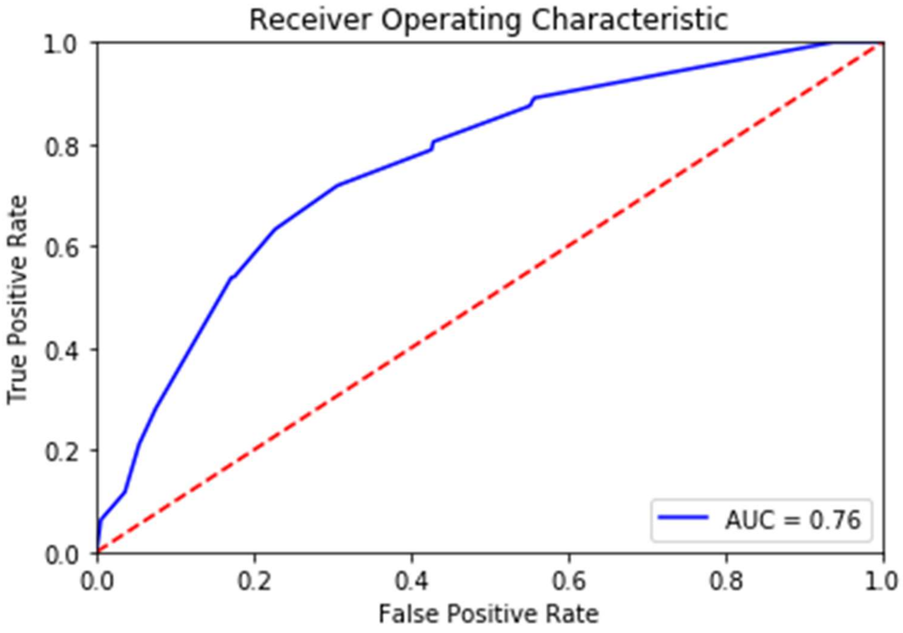
Confusion Matrix

Matrix confirms that out 1071 test records, 929 (912+17) records (86.74%) have been classified correctly.

True Positive : 912 records Annual Income: Below Average Classified as : Below Average Income	**False Negative** : 31 records Annual Income: Above Average Classified as : Below Average Income
False Positive : 111 records Annual Income: Below Average Classified as : Above Average Income	**True Negative** : 17 records Annual Income: Above Average Classified as : Above Average Income

ROC / AUC Curve

AUC curve at 76% shows that classes are separated with certain overlap.



Decision Tree

Gini Index criteria with maximum 15 leaf nodes are used to perform classification. Post applying criteria on train data and validating the test data model predicts accuracy of 88.32%.

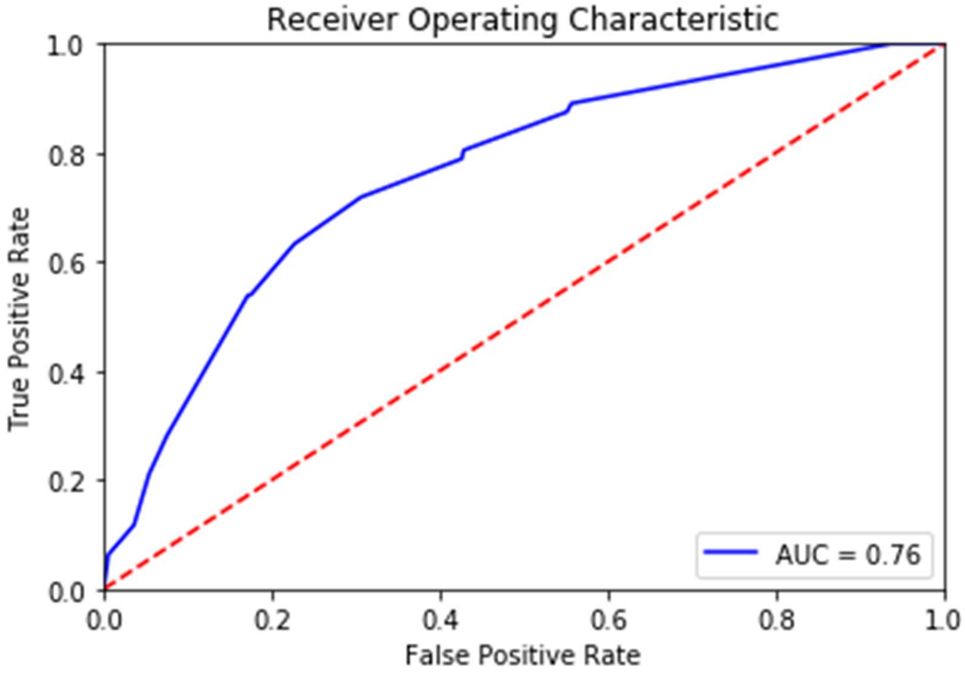
Confusion Matrix

Matrix confirms that out 1071 test records, 946(938+8) records (88.32%) have been classified correctly.

True Positive : 938 records Annual Income: Below Average Classified as : Below Average Income	**False Negative** : 5 records Annual Income: Above Average Classified as : Below Average Income
False Positive : 120 records Annual Income: Below Average Classified as : Above Average Income	**True Negative** : 8 records Annual Income: Above Average Classified as : Above Average Income

ROC / AUC Curve

AUC curve at 76% shows that classes are separated with certain overlap.



Logistic Regression

LBFGS solver is used to perform Logistic Regression using which model provides accuracy of 88.14%. ‵liblinear’,’sag’, ’saga’ solvers are also used but all the solves are providing almost same accuracy.

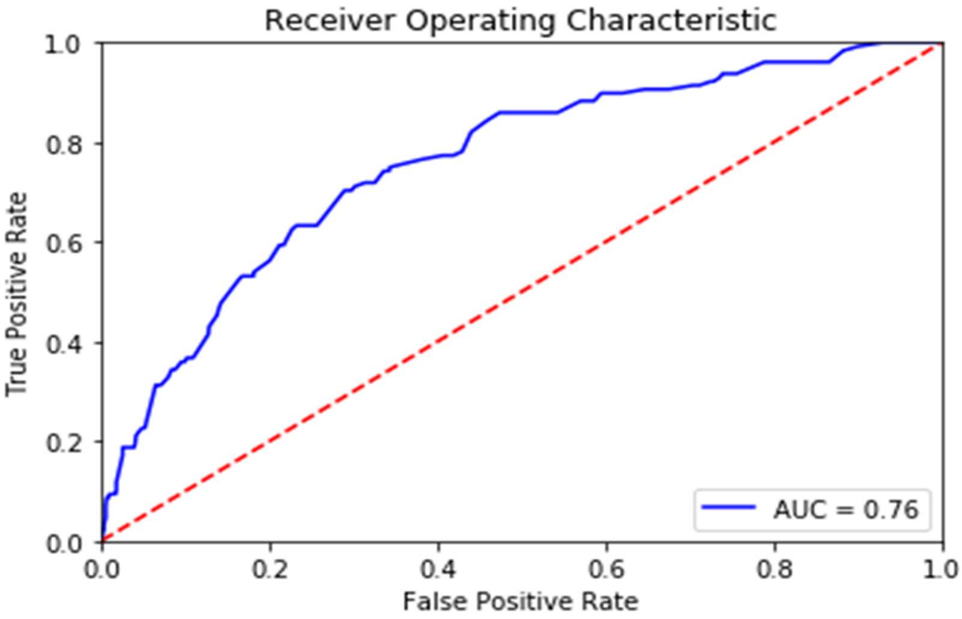
Confusion Matrix

Matrix confirms that out 1071 test records, 944 (940+4) records (88.14%) have been classified correctly.

True Positive : 940 records Annual Income: Below Average Classified as : Below Average Income	**False Negative** : 3 records Annual Income: Above Average Classified as : Below Average Income
False Positive : 124 records Annual Income: Below Average Classified as : Above Average Income	**True Negative** : 4 records Annual Income: Above Average Classified as : Above Average Income

ROC / AUC Curve

AUC curve at 76% shows that classes are separated with certain overlap.



KNN

KNeighbours classifier is the only model amongst all the 6 models that provides lowest accuracy of 85.15 % along with lowest AUC of 68 %.

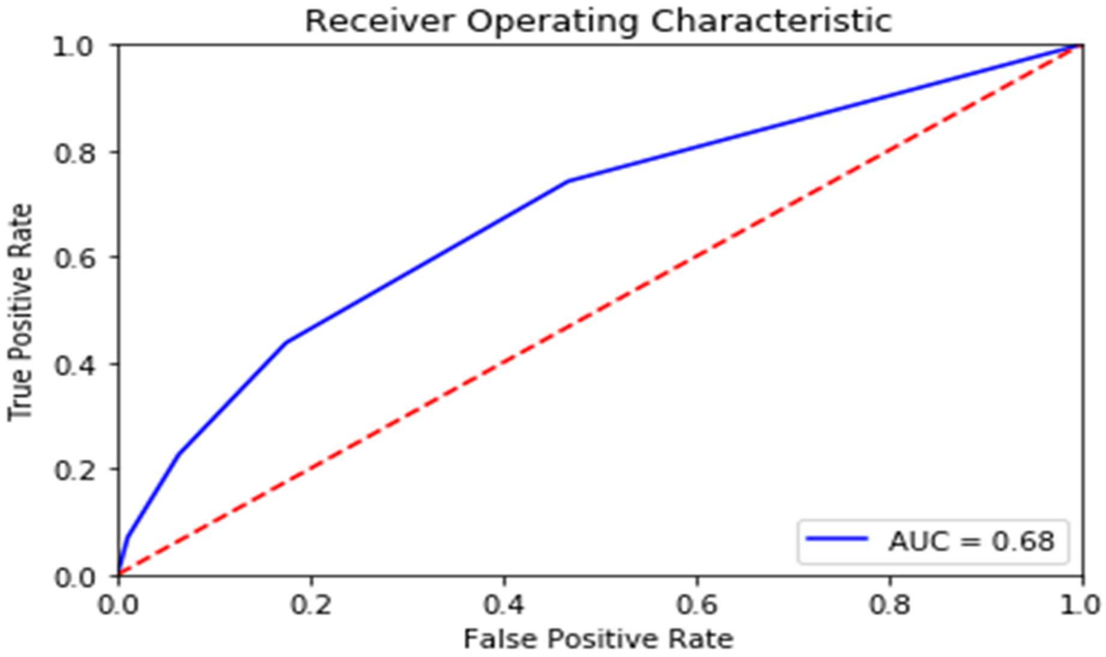
Confusion Matrix

Matrix confirms that out 1071 test records, 912 (883+29) records (85.15%) have been classified correctly.

True Positive : 883 records Annual Income: Below Average Classified as : Below Average Income	**False Negative** : 60 records Annual Income: Above Average Classified as : Below Average Income
False Positive : 99 records Annual Income: Below Average Classified as : Above Average Income	**True Negative** : 29 records Annual Income: Above Average Classified as : Above Average Income

ROC / AUC Curve

With close to 160 records being misclassified, AUC curve stands at 68%.



Support Vector Machine

Linear model is used to perform SVM classification. Post validating test data, model predicts accuracy of 88.04%.

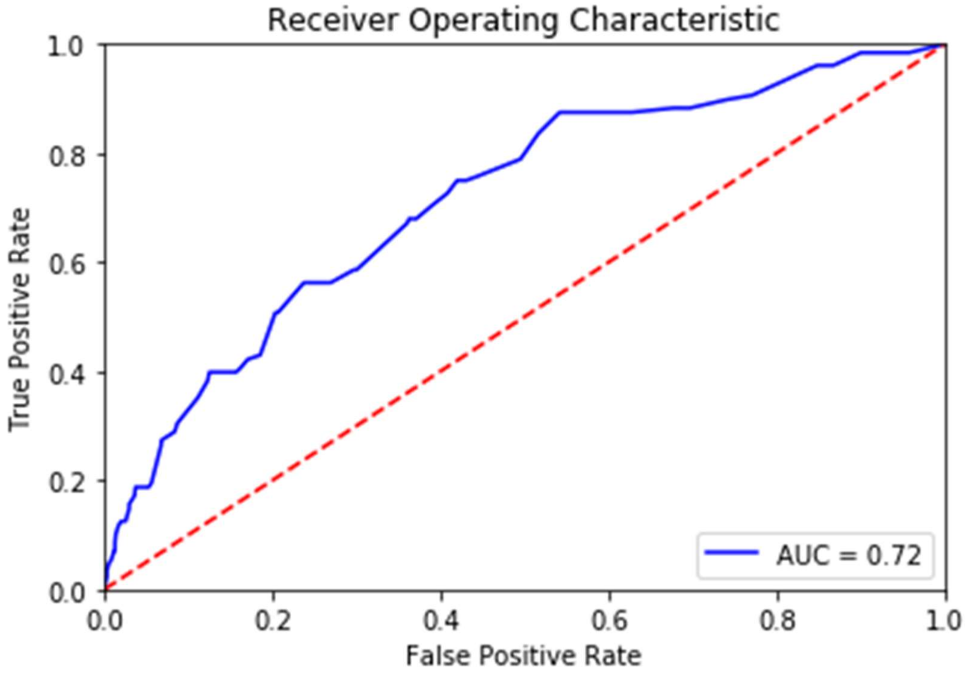
Confusion Matrix

Matrix confirms that out 1071 test records, 943 records (88.14%) have been classified correctly.

True Positive : 943 records Annual Income: Below Average Classified as : Below Average Income	**False Negative** : 0 records Annual Income: Above Average Classified as : Below Average Income
False Positive : 128 records Annual Income: Below Average Classified as : Above Average Income	**True Negative** : 0 records Annual Income: Above Average Classified as : Above Average Income

ROC / AUC Curve

AUC curve at 72% shows that classes are separated with certain overlap.



Random Forest

Using RandomForestClassifier model predicts accuracy of 88.23%.

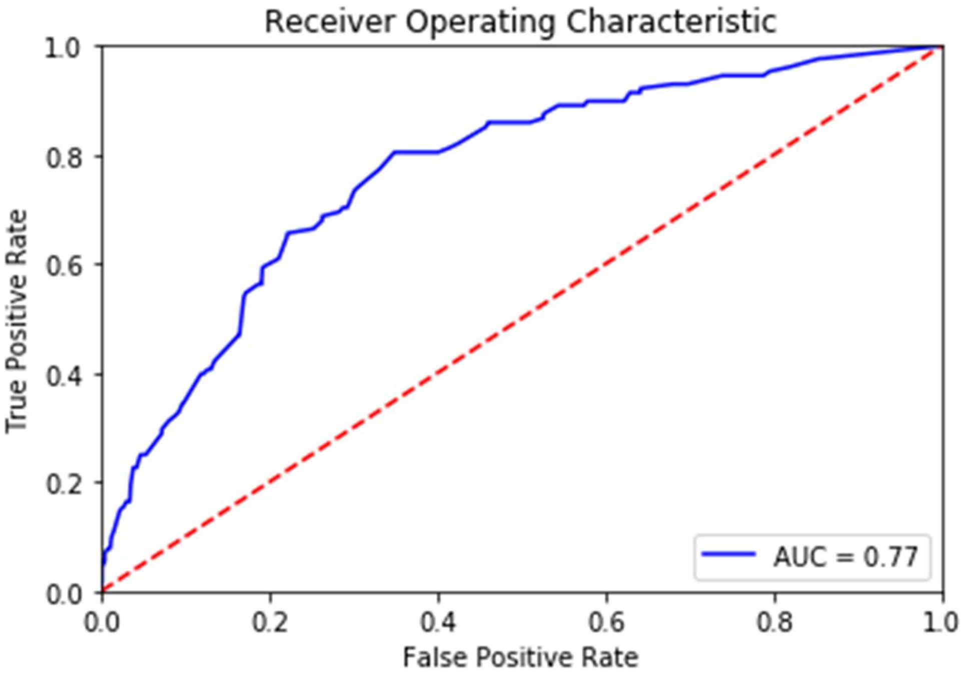
Confusion Matrix

Matrix confirms that out of 1071 records 945 (935+10) records (88.23%) have been classified correctly.

True Positive : 935 records Annual Income: Below Average Classified as : Below Average Income	**False Negative** : 08 records Annual Income: Above Average Classified as : Below Average Income
False Positive : 118 records Annual Income: Below Average Classified as : Above Average Income	**True Negative** : 10 records Annual Income: Above Average Classified as : Above Average Income

ROC / AUC Curve

AUC curve at 77% shows that classes are separated with certain overlap.



○ Identification of best model backed by logic for selecting it

Summarized view of all 6 models applied shows that Prediction Accuracy and AUC Ratio is more or less same in all the model.

Model Name	Model Accuracy	AUC Ratio	No of Record (Correctly Classified)
Naïve Bayes	86.74%	76%	929
Decision Tree	88.32%	76%	946
Logistic Regression	88.14%	76%	944
KNN	85.15%	68%	912
Random Forest	88.23%	77%	945
SVM	88.14%	72%	943

Though SVM model has prediction accuracy of 88.04 % but has AUC ratio of only 72% with 0 records classified as False Positive and True Negative.

Random Forest, KNN, Logistic Regression, Decision Tree all show nearby same accuracy and AUC Ratio but amongst all the model Logistic Regression has second highest number of correctly classified records but better AUC ratio, hence Logistic Regression Model is chosen as best model amongst all the models.

○ Lessons learnt from project

Understanding of data from both the perspective i.e. technical and functional (Domain) is very important to proceed ahead with data cleaning and exploration. It is the base of constructing a good classification model.

Data analysis and cleaning is an important task as at this stage lot of important decision are to be made i.e. whether to replace blank data with NaN value or Mean value of rest of the data. Data analysis helps to decide kind of encoding to be used for conversion of categorical data to numerical.

Data exploration helps to visualize the relationship between predictor and response variable. It helps to identify highly correlated variable which can be attributed to train / test data.

Any kind of misjudgement either at Data analysis / cleaning or Data Exploration may lead incorrect output in terms of Model Prediction Accuracy or Confusion Matrix which may ultimately lead to revisit the entire approach beginning from stage 1 i.e. Understanding the data.

More the number of techniques, better the chances to find anomalies or error. Merely applying only technique will not give best model hence it is advisable to use all the techniques and compare the results to find out best of model for a given data set.