Solution:

**A.Exploratory Data Analysis:**

Data shape: (40935, 15)

1. **Data types of columns:**

|  |  |
| --- | --- |
| **Variable** | **Datatype** |
| Age | float64 |
| WorkClass | object |
| fnlwgt | float64 |
| Education | object |
| EducationNum | float64 |
| MaritalStatus | object |
| Occupation | object |
| Relationship | object |
| Gender | object |
| CapitalGain | float64 |
| CapitalLoss | float64 |
| HoursPerWeek | float64 |
| NativeCountry | object |
| Income | object |

**2: Descriptive Stats for numerical columns**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Unnamed: 0** | **Age** | **fnlwgt** | **EducationNum** | **CapitalGain** | **CapitalLoss** | **HoursPerWeek** |
| **count** | 40935 | 38194 | 38149 | 38132 | 38129 | 38217 | 38255 |
| **mean** | 14002.28052 | 38.59158507 | 189804.2975 | 10.08016889 | 1076.079467 | 88.53693383 | 40.40713632 |
| **std** | 9133.039458 | 13.68274859 | 105929.6002 | 2.570247353 | 7440.867512 | 406.1040969 | 12.43282347 |
| **min** | 0 | 17 | 12285 | 1 | 0 | 0 | 1 |
| **25%** | 6352.5 | 28 | 117444 | 9 | 0 | 0 | 40 |
| **50%** | 12711 | 37 | 178251 | 10 | 0 | 0 | 40 |
| **75%** | 21277.5 | 48 | 237868 | 12 | 0 | 0 | 45 |
| **max** | 32560 | 90 | 1490400 | 16 | 99999 | 4356 | 99 |

**3.Null Values Count in each column**

|  |  |
| --- | --- |
| **Variable** | **Null Count** |
| Age | 2741 |
| WorkClass | 2789 |
| fnlwgt | 2786 |
| Education | 2743 |
| EducationNum | 2803 |
| MaritalStatus | 2924 |
| Occupation | 2759 |
| Relationship | 2815 |
| Gender | 2727 |
| CapitalGain | 2806 |
| CapitalLoss | 2718 |
| HoursPerWeek | 2680 |
| NativeCountry | 2729 |

**4.Unique Values in categorical Columns**

**1.WorkClass**

|  |  |
| --- | --- |
| Private | 26529 |
| Self-emp-not-inc | 2962 |
| Local-gov | 2443 |
| ? | 2222 |
| State-gov | 1544 |
| Self-emp-inc | 1314 |
| Federal-gov | 1107 |
| Without-pay | 16 |
| Never-worked | 9 |

2.Education

|  |  |
| --- | --- |
| HS-grad | 12356 |
| Some-college | 8501 |
| Bachelors | 6243 |
| Masters | 2102 |
| Assoc-voc | 1622 |
| 11th | 1444 |
| Assoc-acdm | 1260 |
| 10th | 1079 |
| 7th-8th | 735 |
| Prof-school | 646 |
| 9th | 598 |
| 12th | 501 |
| Doctorate | 464 |
| 5th-6th | 382 |
| 1st-4th | 195 |
| Preschool | 64 |

**3.EducationNum**

|  |  |
| --- | --- |
| 9 | 12252 |
| 10 | 8533 |
| 13 | 6281 |
| 14 | 2080 |
| 11 | 1610 |
| 7 | 1427 |
| 12 | 1259 |
| 6 | 1088 |
| 4 | 740 |
| 15 | 641 |
| 5 | 607 |
| 8 | 505 |
| 16 | 463 |
| 3 | 386 |
| 2 | 195 |
| 1 | 65 |

**4.MaritalStatus**

|  |  |
| --- | --- |
| Married-civ-spouse | 17342 |
| Never-married | 12664 |
| Divorced | 5101 |
| Widowed | 1207 |
| Separated | 1195 |
| Married-spouse-absent | 473 |
| Married-AF-spouse | 29 |

**5.Relationship**

|  |  |
| --- | --- |
| Husband | 15336 |
| Not-in-family | 9885 |
| Own-child | 5962 |
| Unmarried | 3950 |
| Wife | 1804 |
| Other-relative | 1183 |

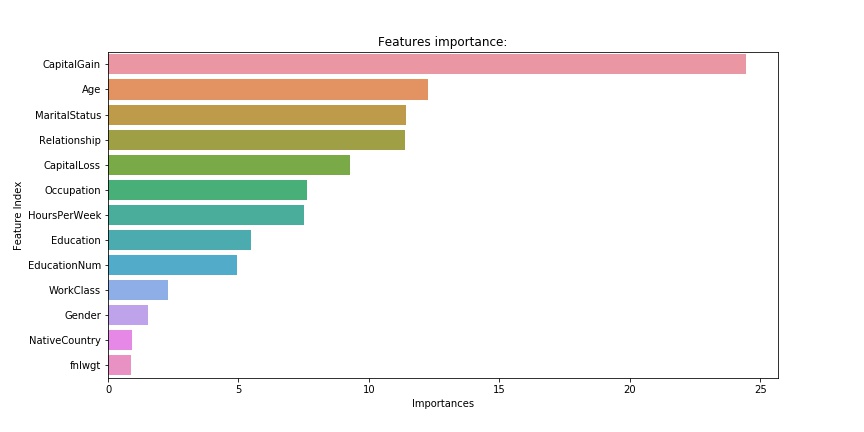
**6.Gender**

|  |  |
| --- | --- |
| Male | 25506 |
| Female | 12702 |

**7.Native Country**

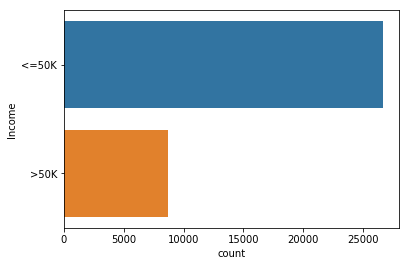
|  |  |
| --- | --- |
| United-States | 34265 |
| Mexico | 744 |
| ? | 667 |
| Philippines | 234 |
| Germany | 153 |
| Puerto-Rico | 144 |
| Canada | 142 |
| India | 127 |
| El-Salvador | 118 |
| Cuba | 109 |
| China | 106 |
| England | 99 |
| Italy | 85 |
| Jamaica | 84 |
| Dominican-Republic | 83 |
| South | 82 |
| Japan | 75 |
| Vietnam | 74 |
| Poland | 70 |
| Guatemala | 65 |
| Columbia | 64 |
| Haiti | 58 |
| Portugal | 54 |
| Taiwan | 53 |
| Iran | 46 |
| Nicaragua | 41 |
| Greece | 39 |
| Peru | 36 |
| France | 29 |
| Ireland | 29 |
| Ecuador | 28 |
| Thailand | 27 |
| Cambodia | 23 |
| Hong | 23 |
| Outlying-US(Guam-USVI-etc) | 22 |
| Trinadad&Tobago | 21 |
| Yugoslavia | 19 |
| Laos | 19 |
| Scotland | 17 |
| Hungary | 16 |
| Honduras | 15 |
| Holand-Netherlands | 1 |

**Feature Importance**



**Insights**

Class imabalnce of target variable is evident . 75% of data belongs to <=50k category whereas only 25% data belongs to >50k category



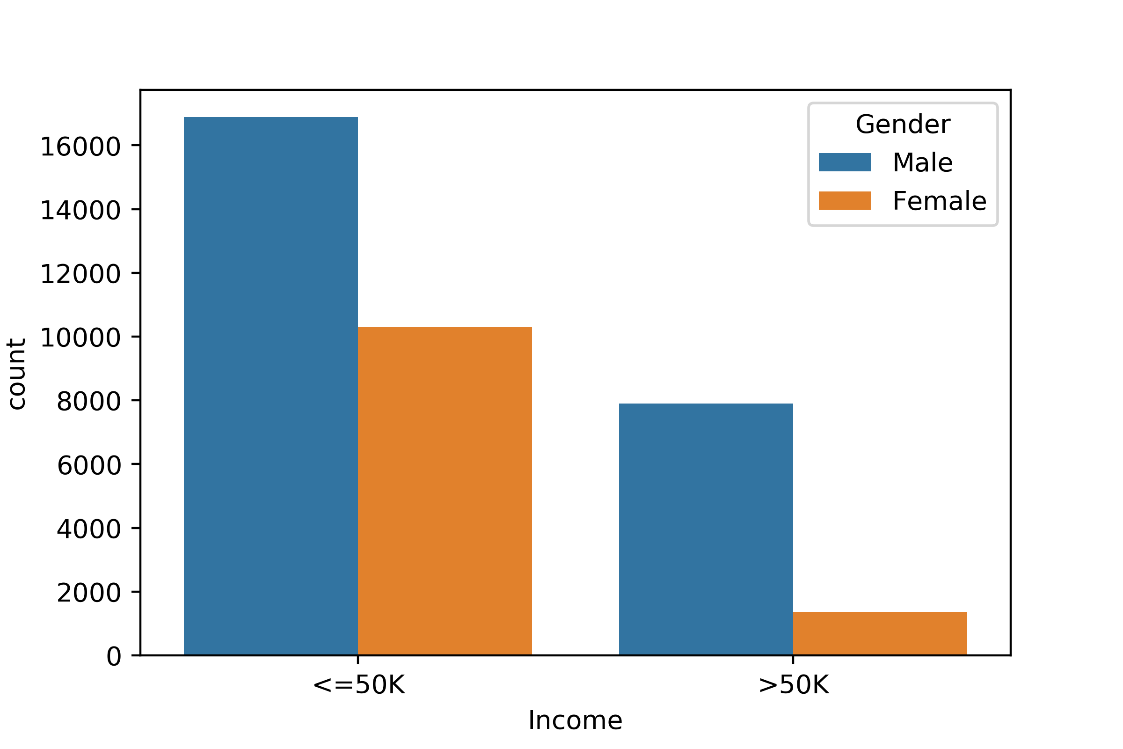
Top 3 Average Hours per week beings spent by people in Farming Fishing (47 ),Exec managerial and Transport moving (45) respectively .Lowest 2 are Priv house service and Other-Service.

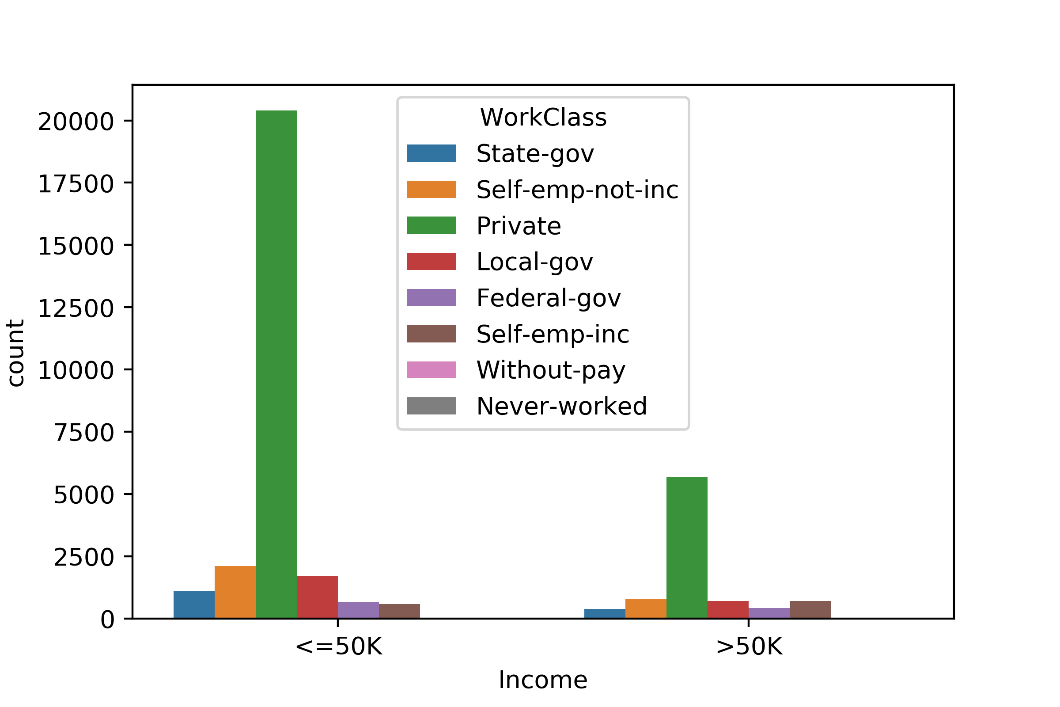
Insight:Top 3Count of people all over the countries in the data with educations with Income>50K as well as income <=50k are Bachelors,HS-Grad and some- college.

Insight:Top 3 countries with 3 Highest Education levels (Maters,Doctrate and Prof-School )are US ,India and China

|  |  |
| --- | --- |
| Countries | Count |
| United-States | 2559 |
| India | 50 |
| China | 33 |

Insight: Male is dominating in both the classes of income as there number is high all over the world. Very marginal proportion of Female can be seen in >50k income group.





Insight:Private workclass dominates in both the income group followed by Self Employee not inc and local gov

**B. Model Performance**

**Confusion Matrix:**

|  |  |  |  |
| --- | --- | --- | --- |
| Actual | Predicted | | |
|  | <=50K | >50K |
| <=50K | 7358 | 442 |
| >50K | 940 | 1494 |

TN(True Negatives)=Out of 7800 cases of <=50K income person model was able to correctly identify 7358

TP(True Positives)=Out of 2434 cases of >50k income person model was able to correctly predict 1494

Out of all 10234 cases of Income test data ,classifier predicted correctly 8852

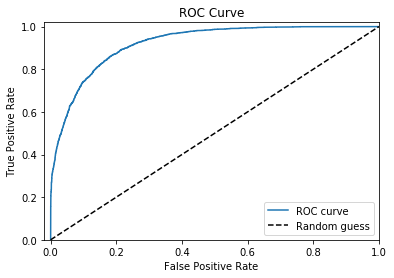
Accuracy=8852/10234\*100=86.4%

Precision=TP/TP+FP=0.77

Recall=TP/TP+FN=0.61

AUC: 0.92

ROC Curve:



**Feature Importance**