EMPLOYEE INCOME PREDICTION

Objective: To predict whether a person is earning more than $50K or not

Data: 40935 observations & 14 columns including dependent variable

7 categorical & 6 numerical variables

Personal demographic information like education, occupation, marital status etc.

Every variable has around 7% missing values

Step: 1. Initial univariate analysis

2. Outlier Capping

3. Rule wise missing Treatment

4. Bivariate analysis

5. Feature engineering

6. 70:30 train & test data splitting

6. Model building

7. Model validation

Observations from data:

1. Around 65 % of people are employed in private sector & more than 12% people have government jobs
2. Around 65 % of people have educational qualification of high school or more
3. Every 1 out of 5th person has degree required for sophisticated jobs
4. Around 42% people are married and live with their partner and rest don't have partners
5. Males are twice in proportion compared to females
6. More than 80% people are American native
7. In general, immigrants are more likely to be earning salary greater than $50K
8. There is a good representation w.r.t. occupation and most people are part of skilled labour force
9. Highly educated people have higher capital gain
10. Highly educated people have lower capital loss
11. People working in gov jobs have higher chances of getting paid more than $50K
12. People having advanced degrees have higher chances of getting paid more than $50K

13. Around every 1 out of 2 married person earns salary more than $50K
14. People working at managerial or doing speciality services are more likely to earn salary greater than $50K
15. Males are 3 times more likely to be paid salary more than $50K

Model Result:

1. **Decision Tree:**
   Train accuracy: 83.9%
   Test accuracy: 83.7%

2. **Random Forest:**
   Train accuracy: 85.8%
   Test accuracy: 85.2%

   Train AUC: 0.885
   Test AUC: 0.876

Top 5 important variables (In decreasing order):

CapitalGain
Relationship_husb_wife_Flag
Married_Flag
EducationNum
Degree_Flag