**Prediction of Employees Salary Based on US Census Data**

**Group7** :

**Student Name Roll Number**

Shreya Pal 201414

Arijit Gosh 201276

Sangita Lingesh Rathod 201402

1. **Problem Desciption:-**

We have US Census Data of year 1994. The Dataset was in Excel having 13 features and 31,978 rows. Our task is to analyse the data and predict the salary of employees whether they earn more than 50k dollars or not.

1. **Data Description**:-

There are 4 integer type features :

1. Age - age of an employee varies from 17yrs - 90yrs.
2. Capitalgain - gain that an employee makes to his company ,varies from 0-99,999 .
3. Capitalloss - loss that an employee makes to his company , varies from 0-2415.
4. Hoursperweek- duration of work per week, varies from 1 to 99 hours.

There are 9 categorical features:

1. workclass : a general term to represent the employment status of an individual. Private,Self-emp-not-inc,Local-gov,State-gov,Self-emp-inc,Federal-gov,Without-pay.
2. education : The highest level of education achieved by an individual.HS-grad,Some-college,Bachelors,Masters,Assoc-voc,11th,Assoc-acdm,10th,7th-8th,Prof-school,9th,12th,Doctorate,5th-6th,1st-4th,Preschool.
3. maritalstatus : marital status of an individual.Married-civ-spouse,Never-married,Divorced,Separated,Widowed,Married-spouse-absent,Married-AF-spouse
4. occupation :Prof-specialty,Craft-repair,Exec-managerial,Adm-clerical,Sales,Other-service,Machine-op-inspct,Transport-moving,Handlers-cleaners,Farming-fishing,Tech-support,Protective-serv,Priv-house-serv,Armed-Forces.
5. relationship: Husband,Not-in-family,Own-child,Unmarried,Wife,Other-relative.
6. race: White,Black,Asian-Pac-Islander,Amer-Indian-Eskimo,Other.
7. sex: Male,Female.
8. nativecountry:United-States,Mexico,Philippines,Germany,Puerto-Rico,Canada, El-Salvador,Cuba,England,Jamaica,South,Italy,China,Dominican-Republic,Vietnam,Guatemala,Japan,Poland,Columbia,Iran,Haiti,Taiwan,Portugal,Nicaragua,Peru,Greece,FrancEcuador,Ireland,Hong,Cambodia,Trinadad&Tobago,Laos,Thailand,Yugoslavia,Outlying-US(Guam-USVI-etc),Hungary,Honduras,Scotland,Holand-Netherlands.
9. Target Variable : Over50k- Salary of employee whether 50k above or below
10. **DATA PREPROCESSING**

**Data cleaning**:- There is no null values in our dataset but we can see some special characters “?” in our dataset . So we have identified and replaced those values by stratified sampling method. (in the proportion of the classes).

From output we observe that the features ”workclass”(1809 has ‘?’) and “occupation” (1816 has ‘?’).

[changed target column class names : “<=50k” to “lessthan50k” and “>50k “ to “more than50k” for our convinience ].

Handling Imbalanced Data:- Over50k Salary is approx. 76% while below50k Salary is 24%. We will balance it using SMOTE Method. After balancing the data we got 48,566 rows with lessthan50k =24,283 = morethan50k.

**Handling Outliers**:- we have used 3techniques to find the outliers (we can use anyone of them) :- IQR, Z score, Standard Deviation.

For Age : dataset after removing outliers with iqr:- (48291, 13)

dataset after removing outliers with z\_score:- (48408, 13)

dataset after removing outliers with Standard Deviation: (48408,13)

For Capitalgain : dataset after removing outliers with iqr:- (0,13)

dataset after removing outliers with z\_score:- (48046,13)

dataset after removing outliers with Standard Deviation: (48046,13)

For CapitalLoss: dataset shape after removing outliers with iqr:- (0, 13)

dataset after removing outliers with z\_score:- (45674, 13)

dataset after removing outliers with Standard Deviation: (45674,13)

For hoursperweek: dataset after removing outliers with iqr:- (40088, 13)

dataset after removing outliers with z\_score:- (47803,13)

dataset after removing outliers with Standard Deviation: (47803,13)

After looking at the outliers we came to the point where we will remove outliers of

* + - age (using z\_score or Standard deviation).
    - hours per week (z\_score or standard deviation).
    - capital gain (average of different groups).
    - capital loss (average of different groups).

For capitalgain:-

* making the group in which we will substitute the mean values of that group and finding the mean of each groups.

'1 - 10001' : 4882.108547251578,

'10001 - 20001': 14254.887837580354,

'20001 - 30001': 24278.90606483336,

'30001 - 40001': 34858.18121000075,

'40001 - 50001': 41310.0,

'50001 - 60001': 53910.134143686344,

'60001 - 70001': 61316.021540074114,

'70001 - 80001': 0,

'80001 - 90001': 86856.17098314573,

'90001 - 100001': 99999.0,

'100001 – 110001’: 0

We will replace those values of capitalgain which ranges from 1-10001 to its mean value 4882.1085 and so on

* capital-gain after handling outliers:

0.000000 : 41079

4882.108547 : 4406

14254.887838 : 1487

99999.000000 : 501

24278.906065 : 147

34858.181210 : 5

41310.000000 : 2

61316.021540 : 1

53910.134144 : 1

0.592506 : 1

86856.170983 : 1

0.796228 : 1

**Interpretation**: count of 0 values are 41,079 and so on.

**For CapitalLoss** :

* We do the same thing as for capital gain. Substituting by mean values of that group.

'1 - 1001' : 551.8710057319815,

'1001 - 2001' : 1838.6461536873114,

'2001 – 3001' : 2289.9080840422503,

'3001 - 4001' : 3672.8571428571427,

'4001 - 5001' : 4356.0,

'5001 – 6001' : 0

* Capitalloss after handling outliers

0.000000 : 44587

1838.646154 : 2422

2289.908084 : 558

551.871006 : 57

3672.857143 :7

4356.000000 :1

After handling outliers of all 4 features we have got (47632, 13) rows and columns. We observe that our data is still balanced. For target feature we got

count %

more than 50k 23927 0.50233

less than 50k 23705 0.49767

Visualization of each features we observe that:

* We can group the age columns into bins.
* For capital Gain and Capital Loss, the data is highly skewed which needs to be tackled.
* The hours per week can also be split into bins.

**Analysis of each feature**:

1. **Sex** :- There are more male employees than female employees. From the bivariate histogram plot : there are more male employees having salary morethan50k dollars. Among females, lessthan50k are more in number.
2. **Workclass**:- 72% of employees belong to private workclass category. From bivariate histogram plot: Among private sector there are more employees earning lessthan50k dollars.
3. **Race**:- 87% of total employees are of White race. Majority of employees are of White ,Black and Asian-Pac-Islander races while all others are very few in count. So we will combine all other race into one class labeled as “Others”.
   1. White 41685
   2. Black 3605
   3. Asian-Pac-Islander 1485
   4. others 857

#### Age:- The violen plot shows maximum number of employees who earns "less than 50k" dollars have age approx 25 and who earn "more than 50k" dollars have age approx 45.

* 1. Creating Buckets for Age: 0-25 ,25-50, 50-75, 75-100 as ‘young’, ‘ adult’, ‘senior’, ‘old’.
  2. From bivariate histogram :- There are more adults in our dataset and very less number of old people.

1. **Hoursperweek:**- creating buckets for hoursperweek: 0-40,40-80 as ‘Less hours’ and ‘Normal hours’. There are 61% of employees belong to ‘Less hours’ class. From bivariate histogram :- people who work ‘Lesshours’ tend to have salary more than 50k dollars.
2. **Marital Status**:-
   1. Married-civ-spouse 59.856819
   2. Never-married 23.482113
   3. Divorced 10.864545
   4. Separated 2.303074
   5. Widowed 2.015452
   6. Married-spouse-absent 1.137890
   7. Married-AF-spouse 0.340107
3. **Relationship:**
   1. Husband 53.466157
   2. Not-in-family 19.684246
   3. Own-child 10.600017
   4. Unmarried 7.874958
   5. Wife 6.254199
   6. Other-relative 2.120423
   7. The feature ‘Relationship’ and ‘maritalstatus’ have no missing values and there is no complete overlap between the classes of these two features. The overlap is only between ‘husband’ , ‘wife’ of ‘‘Relationship’’ feature and ‘married’ of ‘‘MaritalStatus’’. So we will keep both these features.
4. **Occupation** :-the distribution of income varies across all various occupations. The categories are already uniquely identificable and we will keep them as it is.
5. **Education**:- we will combine school students from ‘preschool’ to 12th
   1. into one class labeled as “no college/university” as count of school students earning more than 50k dollars will be very less.
6. **Native country**:-91% of employees comes from US . Thus we make US as one class and rest countries labeled as “others”.
7. **Capitalgain and capitalLoss**:- approx. 86% of employees have capitalgain as 0 dollars and 93% of employees have capital loss 0 dollars.
   1. Using feature generation , as ‘capitalLoss’ and ‘capitalgain’ are related we drop both columns and their difference is labeled as ‘Capital Difference’. For analyzing the data we created 2 buckets :”Minor” for -5000 dollars to 5000 dollars and “Major” for 5000 dollars to 100,000 dollars.
   2. Capital difference= capital gain + capital loss.
   3. 95% of employees belong to minor class of “capital difference” feature. In minor category there are more number of employees having salary greater than 50k dollars.
   4. [note :We have converted all 4 numerical features into categorical by creating buckets]
   5. Chi square test for independence between each categorical feature and target variable is also computed. All features are dependent on target variable.
   6. We now have 12 columns.
8. **Salary** :- we have assigned “lessthan50k” to 0 and “morethan50k” to 1.
9. **Model Building**

We have split the original dataset into train(60%), validation(20%) and test set (20%)

**Input and Target Columns:**

Separating train, validation and test dataset into 2 parts Target column and other features.

Using One hot encoder we change categorical data into binary.

Separate each class of the binary feature into arrays and then print names of all binary features along with their class names as:

['workclass\_ Federal-gov', 'workclass\_ Local-gov',…. 'education\_ Assoc-acdm', 'education\_ Assoc-voc', 'education\_ Bachelors', 'education\_ Doctorate',…]

Combining these binary features of test, train & validation set with 11 original features , we now have 71 columns for each of the sets.

1. **Training a Logistic Regression model**

We take linear combination (or weighted sum of input features)and apply Sigmoid function (f(z)=1/1+e^z)to the result to obtain a number between 0 and 1. This number represents the probability of the input being classified as “Yes” and we used Confusion matrix to evaluate results.

We got regression-coeffient values of all 60 binary features and an intercept.

**Making Predictions and Evaluating the model**:

We got predicted values of train , test and validation set and compared them with the target column. The accuracy of the model on test , validation and train set is above 80%.

Confusion matrix is a useful tool for analyzing how well our classifier can recognize tuples of different classes. (0,0) and (1,1) tell us when the classifer is getting things right , while (0,1) and (1,0) tell us when classifier is getting things wrong.

Interpretation of Training confusion matrix :

79% of employees having salary lessthan50k is correctly predicted.

86% of employees having salary morethan50k is correctly predicted.

14% of employees having salary morethan50k is wrongly predicted.

21% of employees having salary lessthan50k is wrongly precdicted.

Same interpretation goes for test and validation confusion matrix.

1. **Decision Tree Classifer:**

Training and Visualizing Decision Trees:

A decision tree in general parlance represents a hierarchical series of binary decisions.

In logistic regression algorithm we have computed the accuracy of model manually while in decision tree we let the computer figure out the optimal structure & hierarchy of decisions, instead of coming up with criteria manually.

After fitting the training dataset we get predicted target variable .

In predicted target variable of training dataset we get 54% employees having salary morethan50k dollars and 46% having salary lessthan50k dollars.

The training set accuracy is close to 88% . But we can't rely solely on the training set accuracy, we must evaluate the model on the validation set too. We can make predictions and compute accuracy in one step using model.score. We get, 82% accuracy on validation dataset. We got approx. 82.5% accuracy for test dataset.

In predicted target variable of validation dataset , we get 50.28% employees having salary morethan50k dollars and 49.71% having salary lessthan50k dollars.

In predicted target variable of test dataset ,we get 50.37% employees having salary morethan50k dollars and 49.6% having salary lessthan50k dollars.

**Visualization**

We now plot decision tree and find out the Gini value for each binary features.

This is the loss function used by the decision tree to decide which column should be used for splitting the data, and at what point the column should be split. A lower Gini index indicates a better split. A perfect split (only one class on each side) has a Gini index of 0.

**Feature Importance**

Based on the gini index computations, a decision tree assigns a score to input features based on how useful they are at predicting a target variable. Feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node. Node probability= no. of samples that reach the node / total no. of samples. The higher the value the more important the feature.

Here maritalstatus\_Married-civ-spouse is the most important feature followed by hoursperweek\_lesshours.

1. **Gaussian Naïve Bayesian:**

We have made dummy variables for each categorical columns and convert into binary features, since “Sex” feature is already a binary feature. So, we will take it as it is.

We have divided the dataset into train and test dataset in 80:20 and fit the Gaussian naïve Bayes estimator and got accuracy as 79% also find the confusion matrix.

1. **Conclusion:-**

In this project, We have worked on US employees income dataset .We have removed the Outliers, replaced missing values in some features by stratified sampling method, balanced the imbalanced dataset by SMOTE method and analysed each feature one-by-one. Then, we applied some machine learning algorithm like Logistic regression and decision tree classifier on training, validation and test dataset and finding the accuracy of model and finding a confusion matrix for each of datasets training, validation and testing to see the relationship between actual and predicted target variable.

**Key things I learnt**:

1. Correlation matrix tells us about those features having no correlation with target variables and we can remove those features.
2. While there might be no missing value in the dataset, but there can be erroneous data like our dataset has “?” values in some features. We have to handle this by changing “?” value to NaN and then either remove this character or replace it by some other methods.
3. Since the problem is of binary classification problem, we used logistic regression method and get accuracy of model for each dataframe (training, validation and testing) greater then 80%.
4. By using decision tree classifier we draw a decision tree which tells us about the gini index of each binary features of dataframe and getting the feature importance based on gini indices of all binary features.
5. We found the best model is Decision tree with the accuracy 88% followed by Logistic regression with accuracy 82 % and Gaussian naïve bayesian with accuracy 79%.