**Adult Census Dataset**

**Abstract:**

The Adult dataset comes from the Census Bureau, and the aim is to perform classification models to estimate if a given adult earns more than $50,000 annually based on factors like education, hours worked per week, etc

**Introduction:**

\*In this project, we will use a number of different supervised algorithms to precisely predict individuals income using Adult Census Data Set collected from the UCI machine learning repository. We will then choose the best candidate algorithm from preliminary results and further optimize this algorithm to best model the data. Our goal with this implementation is to build a model that accurately predicts whether an individual makes more than $50,000.

**Data Set:** Link to the Data Set - <https://archive.ics.uci.edu/ml/datasets/Adult>

**Data Information:**

**Attributes:** Dataset consists of 14 columns with numerical, categorical and ordinal data types. The dataset's headers are listed below:

Age, Work class, Final weight, Education, Education-Number, Marital-Status, Occupation, Relationship, Race, Sex, Capital-Gain, Capital-Loss, Hours-Per-Week, Native-Country Data Info: The data contains 48842 rows with some missing values(unknown) in Work class (2799 missing values), Occupation (2809 missing values), Native-Country Columns (857 missing values).

**Target:** The target column in our dataset is ‘Adult Barrier’. It has two columns <=50K and >50K.

**Preliminary Analysis:**

\*Before doing any pre-processing steps onto the data, we need to understand the type of data that is available on our dataset.

Table

Description automatically generated

**Fig: Statistics of All Categorical Features in The Dataset**

Table

Description automatically generated

**Let’s use Histograms and see how data is distributed across our dataset.**

Chart, waterfall chart

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

**\***As there are only few duplicate records we can drop them.

Graphical user interface

Description automatically generated with medium confidence

**Data Pre-processing:**

**\***The ‘Workclass’ have around 2799 unknown values which is represented as ‘?’. As the number of unknown values is high, rather than dropping the rows we can replace by ‘Private’, which have more number of rows than compared to others.

**\***The ‘Native Country’ have around 857 unknown values , which is represented as ‘?’.Similarly we can replace by ‘United-States’, which have more number of rows than compared to others.

**\***Similarly,‘Occupation’ column have around 2809 unknown values ,which is represented as ‘?’.As the number of unknown values is high, rather than dropping the row we can replace by ‘Prof-specialty’, which have more number of rows than compared to others.

\*In this dataset when you look at the target variable ‘Adult Barrier’, there are four values which are <=50K ,<=50K. , >50K. , >50K

Text

Description automatically generated

\*So,if the adult barrier is <=50K, <=50K. then it is replaced ‘0’ and

If the adult barrier is >50K,>50K. then it is replaced by ‘1’.

Graphical user interface, application

Description automatically generated

**\***Now, as the target variable have only two distinct values ‘0’ or ‘1’,which will come under Binary Classification Task.

\*Similarly, we replaced row values in Education and Marital-Status Columns as follows

A picture containing scatter chart

Description automatically generated

**Dropping Columns:**

\*In this dataset we can remove ‘Missing Rows’ and ‘Education-Num’ as these columns do not add any value to the dataset.

**Visualization Of Features In The DataSet :**

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

**Relationship of features with target attribute ‘Adult Barrier’:**

Chart, bar chart

Description automatically generated

**\***From the above diagram , we can determine that people in ‘Private’ workclass will get salary more than 50K per annum.

Chart, bar chart

Description automatically generated

\*Here people who are married will get salary more than 50K.

Chart, bar chart

Description automatically generated

\* People with ‘Male’ gender receive income >50K per annum.

\*Also ‘Whites’ receives >50K income.

**Feature Selection Using Chi-Square Test:**

Text

Description automatically generated

\*Calculate p-value using chisquare test. \*Using Hypothesis we frame Null hypothesis [ Independent variable] and alternate Hypothesis [Dependent on target variable].Below code conducts chisquare test on categorical variables and finds categorical features depending on target feature

**Correlation of target attribute with all features :**

Graphical user interface, application, Teams

Description automatically generated

**Modeling and Results:**

**Definition:** Logistic regression is a predictive algorithm using independent variables to predict the dependent variable. It is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables. Mathematically, a logistic regression model has a dependent variable with two possible values. For example, True/false which is represented by an indicator variable where the two values are labeled as 1 and 0.

**Implementation**: Since the prediction we are making in our model is to see if person income is <=50K or >50K, we use logistic regression for this as it is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary).

**Evaluation :**After creating our model, we can evaluate our model using the test data. We are using overall accuracy score, precision, recall to validate the prediction output. Here are we are getting the overall accuracy of 83% which is good to begin with and precision being 70%.

**K – Nearest Neighbor :**

**Definition:** A k-nearest-neighbor algorithm, often abbreviated k-nn, is a statistical method of classifying data that estimates how likely a data point is to be a member of one group or the other depending on what group the data points nearest to it are in. The k-nearest-neighbor is also called as "lazy learner" algorithm.

**Evaluation:** The overall accuracy for k-nn comes out as 83% while for class 1 the accuracy is 68%. But we can still improve on this. Hence, we continue to run our data on other models to check for better accuracy.

**Random Forest:**

**Definition :** The random forest is a classification algorithm consisting of many decision tree.It uses bagging and features randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.The ‘forest’ it builds is an ensemble of decision trees,usually trained with the ‘bagging’ method.

**Evaluation :** We implement this by importing RandomForest Classifier from sklearn.Then running our model metrics given an overall accuracy as 85% and precision is 79%

**Results:**

Now that we have tested our data on various models, we can compare them and select the best model based on the accuracy. Now our evaluation parameters are as follows

1.Overall accuracy: Accuracy is defined as overall correctly predicted data divided by total number of data points present.

2. Precision: Precision is the fraction of relevant instances of output among total number of retrieved instances.

3. Recall: Recall is the fraction of relevant instances that were retrieved.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Accuracy** | **Precision** | **Recall** |
| **K Nearest Neighbor** | 0.833 | 0.68 | 0.58 |
| **Logistic Regression** | 0.838 | 0.70 | 0.56 |
| **Random Forest** | 0.854 | 0.79 | 0.60 |

**\***Overall, we can say Random Forest is the best model to predict whether a person makes over 50K a year.

**Conclusion**

\*As soon as we look at the dataset we realize that this is a US based survey. Mostly people of white and black ethnicity took part in the survey.

\*This dataset contains information about more males than females. This may be because females did not prefer to take the survey. This dataset seems biased to people making <=50K.

\*Most people started working right after their HighSchool. However, some pursue bachelors or masters or doctorate.

\*People who don’t make it to high school earn less than 50K,which is because of lack of skills, education, exposure.

\*Many women earn less than 50K.Its not juts women, but minorities in race also seem to earn less.

\* The minimum age is 17 and the maximum is 90 years, most of the working age group lies between 20-40.