**HDSC Fall’22 Premiere Project Presentation: Adult Income Prediction**

**Introduction**We as a Society, has revolutionized our known fields of expertise drastically. Machine Learning and Artificial intelligence is one of such fields where knowledge is extracted out of raw data. Such knowledge can be used to build dynamic models with high precision and accuracy, such that the model itself can then be used to predict or classify previously unknown data. One of the most essential aspect of providing knowledge from census data is predicting data like income and health of every individual by investigating previous records.

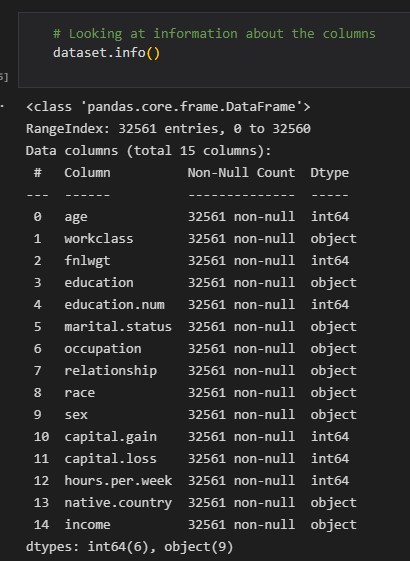
For this project, Team Scrapy used the “[Adult Income Census](https://www.kaggle.com/datasets/uciml/adult-census-income)” dataset from Kaggle. The dataset is credited to Ronny Kohavi and Barry Becker and was drawn from the 1994 United States Census Bureau data. The Team used variables such as age, capital gains or losses, native country, and education level to predict whether an individual will earn more or less than $50,000 per year.

**Data Preparation**

The data preparation phase is where the data was explored for understanding and insights. The data was cleaned, transformed, and integrated. Below is a snippet of the first 5 observations in the data:



It had 14 variables from which to select and use to predict whether the income of the individual will be less or more than 50K a year. The columns in the dataset can be seen in the snapshot below:



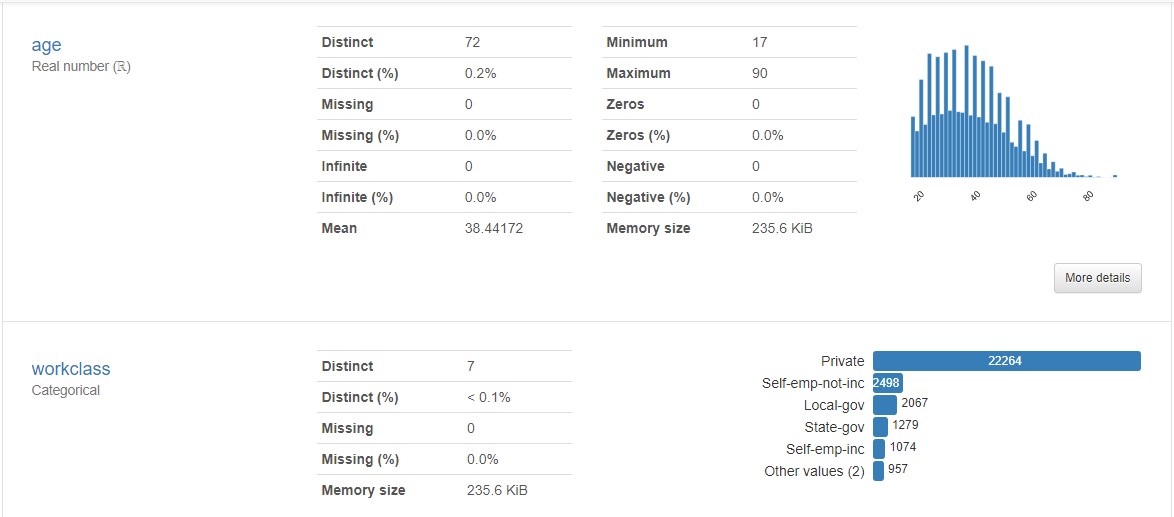
Rows with missing values (represented by “?”) were removed, and the column names restricted to only alphanumeric characters. After necessary data manipulation and cleaning. The dataset had 30,139 rows and 15 columns.

**Exploratory data analysis (EDA)**

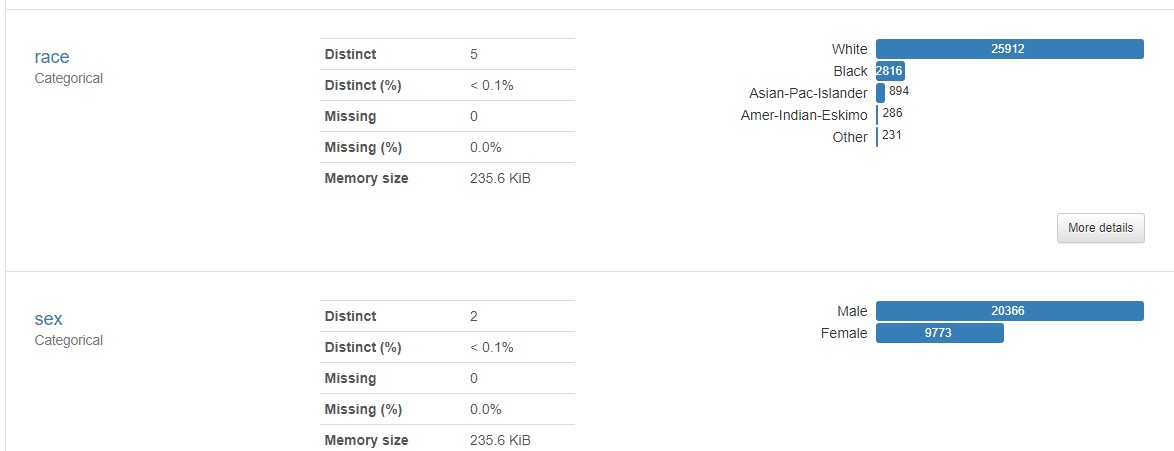
Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

Using the Pandas Profiling library, a report was generated on the dataset, and below are some of the descriptive statistics:

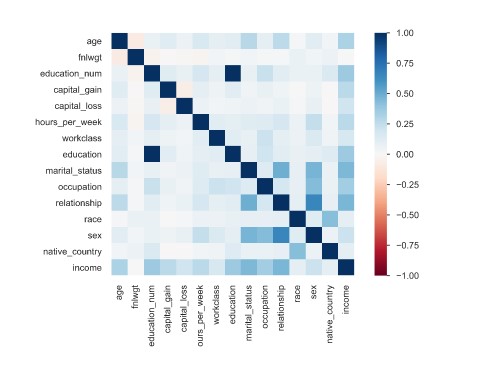
* The youngest and oldest ages in the dataset were 17 and 90 years respectively, with a mean age of 36.44 years
* The work class column contain 7 unique work classes, with individuals working in the private sector representing a significant portion of our dataset.



* 85% of the observations were from white individuals.
* The dataset had twice as many observations on males as it did on females



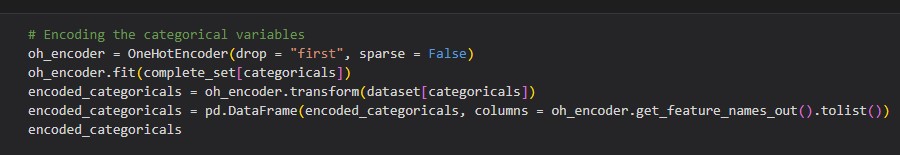
Correlations in the dataset are shown below



**Feature engineering**

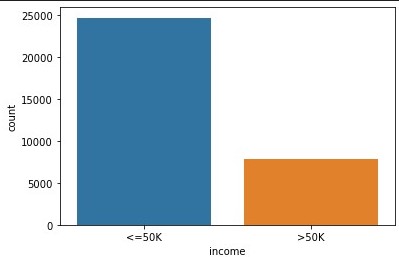
No new features were introduced, but due to the relationship between education and education number, the education column was dropped.

A one-hot encoder was then applied to convert the categorical variables into numeric terms, usable by the models.



After encoding all the categorical features, min-max scaling of the numeric columns was done to mitigate the effects of outliers.

As can be seen below, the dataset was mildly imbalanced, with just about 24% of the observations being for the positive class (above 50k), hence SMOTE was applied to balance the minority.



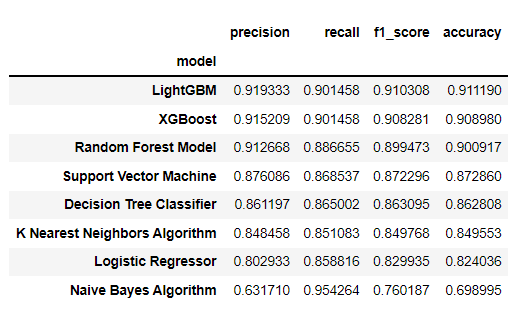
SMOTE (Synthetic Minority Oversampling Technique) is an oversampling technique where the synthetic samples are generated for the minority class. This algorithm helps to overcome the over fitting problem posed by random oversampling. It focuses on the feature space to generate new instances with the help of interpolation between the positive instances that lie together.

**Model training and evaluation.**

The dataset was split into train, test and evaluation sets using the train\_test\_split() function of sklearn and the following classifier models were used:

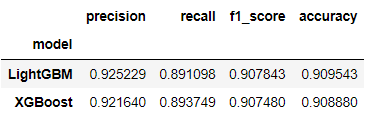
* Logistic regression
* Decision Tree Classifier
* Random forest classifier
* The K-nearest neighbors’ algorithm
* Support vector machines
* Naïve Bayes algorithm
* LightGBM Classifier
* XGBoost Classifier

The performance of the models was examined using accuracy, precision, recall and F1 scores. The results are shown below:



**Observation**

The LightGBM and XGBoost models outperformed the other classifiers, thus they were chosen for evaluation on the unseen data as can be seen below:



The LightGBM model outperformed the XBoost model, and was therefore chosen for further development and use.

**Conclusion**

The aim of today’s research in the field of data science is to build systems and algorithms to extract knowledge from data. The results obtained above can be used as a standard point of reference for other projects done in the field of predicting values from census data. This project can further be used as a basis for improving the present classifiers and techniques resulting in making better technologies for accurately predicting income level of an individual.

**Team Members**

1. Otchere Kwame - **Project Lead**
2. Shrey Gupta - **Assistant Project Lead**
3. Ajisafe Kolade - **Query Analyst**
4. Jacob Agyei
5. Richard Owusu
6. Ayush Kumawat
7. Babarinde Mojeed
8. Nagendranath Vemishetti
9. Adelusi Adegoke
10. Ugo Chukwu
11. Olalekan Adedokun
12. Olaiya Olamibowale