**Introduction:**

The goal of this project is to predict the salary of employees based on various features such as their education, native country, and job position. To achieve this, multiple steps of data handling were implemented along with a few machine learning models which we will be discussing in this report.

**Data Visualization & Analysis:**

The first step in this project was to visualize the data to find the detailed descriptions for every feature of the dataset.

Following the data visualization, we were able to find the underlying issues for each of the features here is what we found:

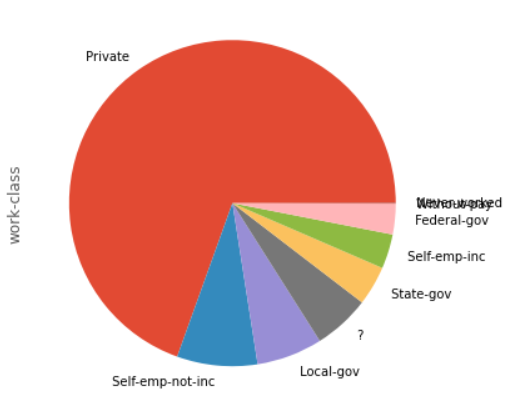


Figure 1 work-class

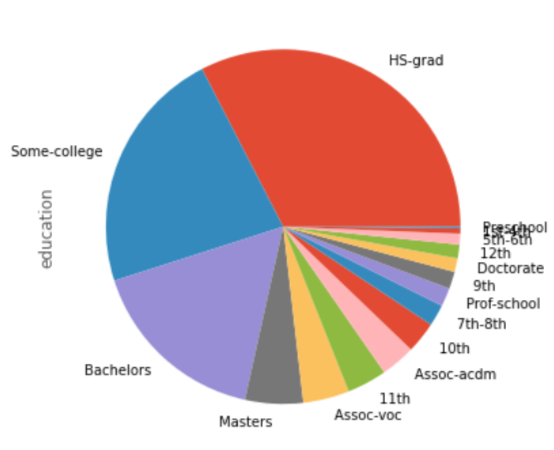
**- work-class:**

Has missing values

Nominal variable

Bias: Private (69.51%)

Figure 2 education



**- Education:**

Ordinal variable

**- Relationship:**

Nominal variable

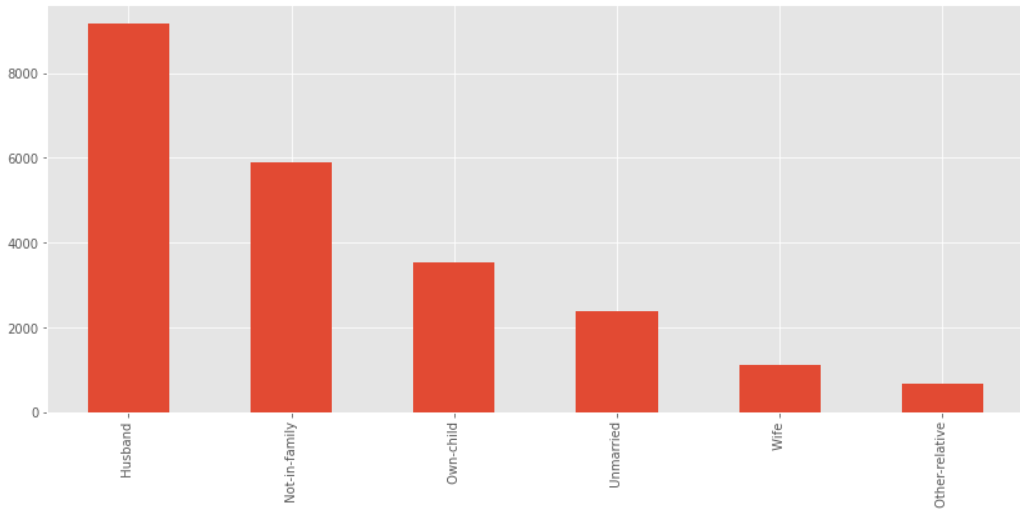


Figure 3 relationship

**- position:**

Has missing values

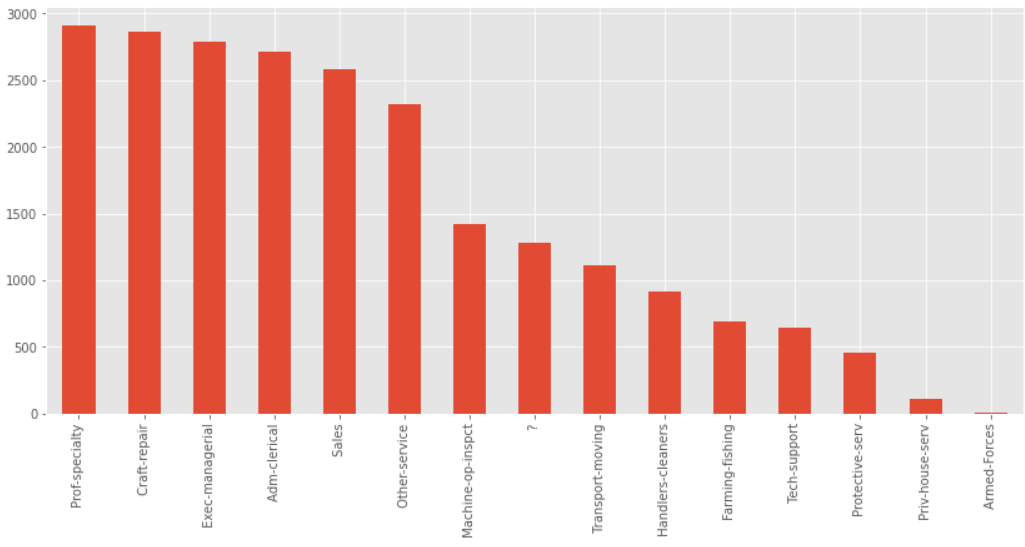


Figure 4 position

Nominal variable

**- race:**

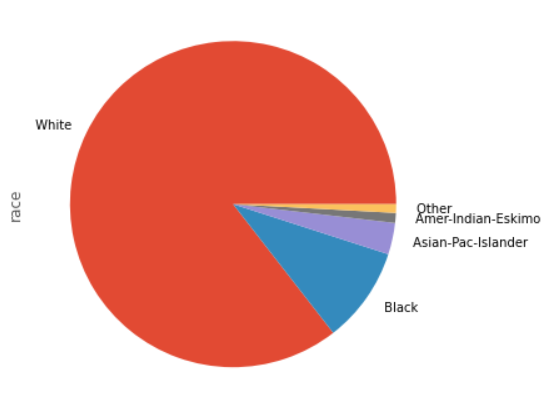


Figure 5 race

Nominal variable

Bias: White (85.53%)

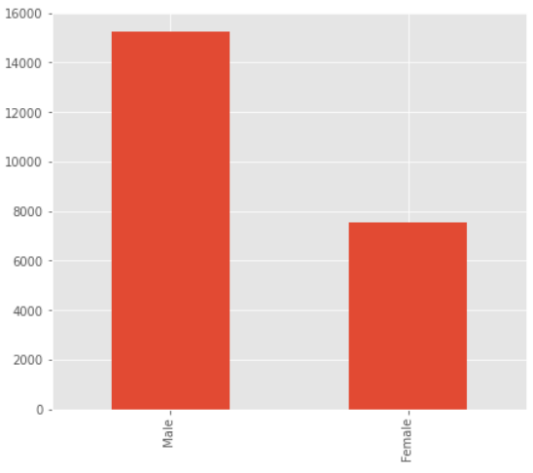


Figure 6 sex

**- sex:**

Nominal variable

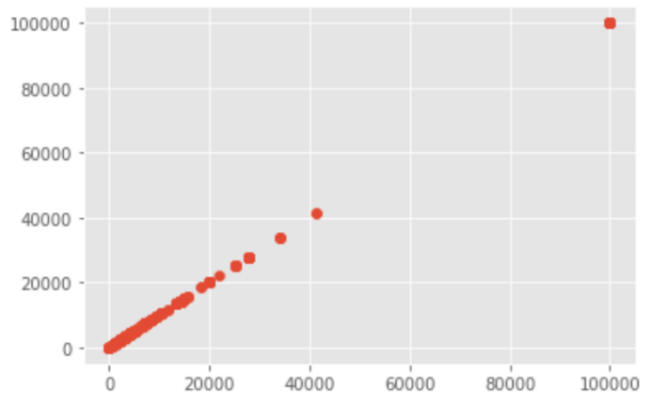


Figure 7 Capital-gain

**- capital-gain:**

Has outliers

**- capital-loss:**

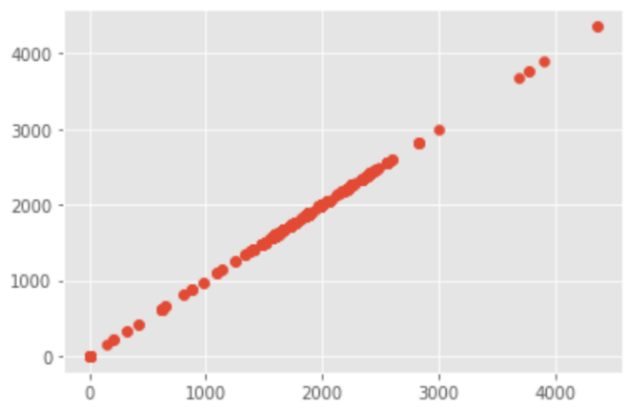


Figure 8 Capital-loss

Has outliers

**- native-country:**

Has missing values

Nominal variable

Bias: United-States (89.57%)

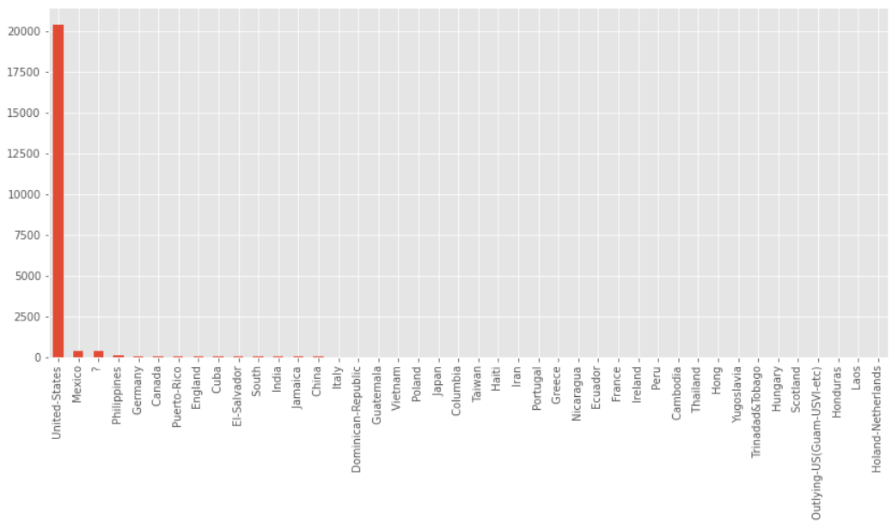


Figure 9 Native-country

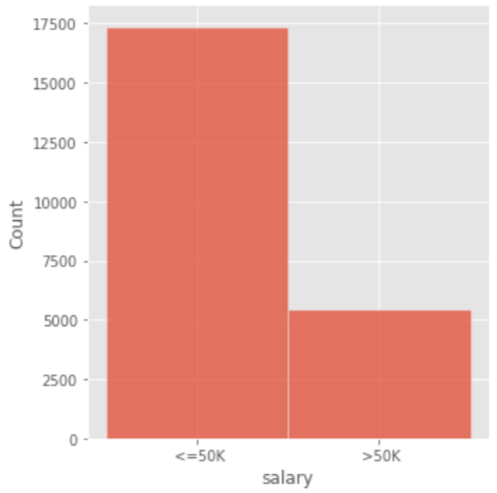


Figure 10 Salary

**- salary:**

Ordinal variable

Bias: <=50K (76.15%)

**Data Preprocessing**

**Replacing missing values:**

In columns work-class, native-country and position, there were missing values. To handle this problem, missing values were filled with the most common value for each column: private, united-states and Exec-managerial respectively.

**Renaming columns:**

Some columns between training and test data sets did not have the same column names so it was necessary to rename these columns for removing inconsistencies and standardizing using the rename function. Workclass, occupation fnlwgt were changed to work-class position work-fnl.

**Dropping columns:**

Columns which were not correlated to the rest of the dataset columns were dropped for faster and improved model training. work-fnl column was dropped for these reasons. We used the drop function for this procedure.

**Encoding categorical data:**

Categorical data needed to be encoded into numerical data to be usable by the model. For that, we used label encoding from sklearn library.

**Training and Testing**

The training dataset has 22792 rows and 15 columns: age, work-class, work-fnl, education, education-num, marital-status, position, relationship, race, sex, capital-gain, capital-loss, hours-per-week, native-country, salary

The testing dataset has 9769 rows and 14 columns: age, work-class, work-fnl, education, education-num, marital-status, position, relationship, race, sex, capital-gain, capital-loss, hours-per-week, native-country.

**Models**

For the classifying models we used five different models which are: MLP, xgboost, GradientBoost, Random Forest and CatBoost.

**MLP:**

Train accuracy: 86.22%

Test accuracy: 84.5%

**xgboost:**

Train accuracy: 86.22%

Test accuracy: 84.5%

**GradientBoost:**

Train accuracy: 86.22%

Test accuracy: 84.5%

**Random Forest:**

Train accuracy: 86.22%

Test accuracy: 84.5%

**CatBoost:**

Train accuracy: 86.22%

Test accuracy: 84.5%