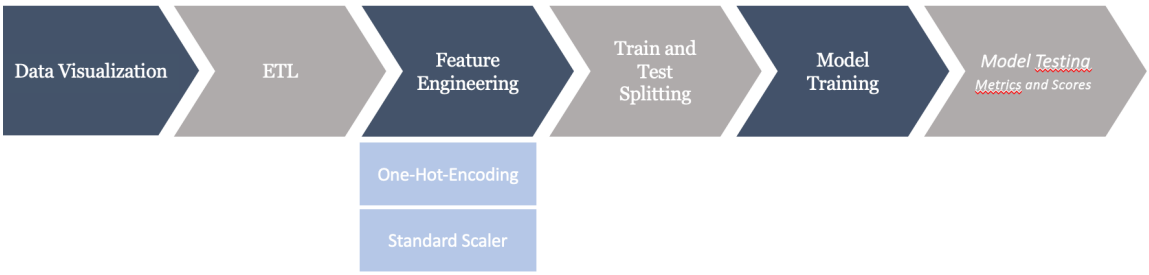
The Lightweight IBM Cloud Garage Method for Data Science



1. Data Source

The Data set is listed under the name “Adult Dataset”, also known as “Census Income Dataset”. It can be found at UCI Machine Learning Repository in the following URL: https://archive.ics.uci.edu/dataset/2/adult

This Data set contains census information and a label column indicating weather an adult income is greater than $50K/yr. The use case that we are going to cover is applying ML techniques to predict this output class based on the census data.

It contains the Following Columns:

- Age
- Workclass
- Fnlwgt
- Education
- Education_num
- Marital_status
- Occupation
- Relationship
- Race
- Sex
- Capital_gain
- Capital_loss
- Hours_per_week
- Native_country
- Wage_class

2. Quality Assessment

During the quality assessment we made sure that the dataset does not contain duplicate values or null values. Since the dataset does not have either null or duplicate values, we can continue with the following steps.

3. Feature Engineering

During Feature Engineering phase, the following actions were taken:

- Since the majority of native_country values are “USA”. We have decided to substitute “USA” as 1 and other countries as “0”. This reduces the number of unique values without
- We have use a LabelEncoder() to the output column “wage_class”. The unique values for this column are “<=50” or “>50”. Applying a LabelEncoder allow as to convert these values to an output labels which are 0 for “<=50” and 1 for “>50”.
- A StandardScaler() have been used for numerical columns, making sure all of them have mean of 0 and standard deviation of 1.
- Finally, a OneHotEncoder() have been used for categorical columns, in order to transform them to numerical values (either 0 or 1) to make them understandable for our ML models.

4. Algorithms used

The use case for this project requires to apply supervised learning algorithms to classify the adults in two different classes:

- Class 0 → Adult income is less or equal than $50K/yr
- Class 1 → Adult income is greater than $50K/yr

We wanted to apply different algorithms to make a comparison an see which one performs better in the task purposed. Also, we want to use traditional ML algorithms and Deep Learning neural networks to see if there is a difference.

The Classification algorithms that we are going to use are the following:

- Gradient Boosting Classifier
- Random Forest Classifier
- XGB Classifier
- Logistic Regression
- K -Nearest Neighbor
- Multi-Layer Perceptron - Keras

As we can see, we are going to apply 5 different traditional ML algorithms and 1 Deep Learning algorithms.

5. Framework

The framework used will be the libraries sckit-learn, XGB and Tensorflow Keras. Since our final deliverable is going to be a Jupyter notebook, no additional frameworks are needed.

6. Model Performance Indicators

Since we are facing a Supervised Learning Classification problem, we are going to use the following metrics:

These metrics allow as to measure how our different models are performing. Specially, we are going to focus on accuracy and F1-Score to see which model performs better and should be chosen.