REFLECTION NOTE – PAULINE ONYANGO

Challenging Financially Inclusive Africa - Zindi Challenge.

Methods Tried

The first step involved **exploratory data analysis (EDA)** to understand the structure of the dataset. The target variable, *bank_account*, was identified, and key features such as age, gender, education level, employment type, and location were examined. A bar chart was used to visualize the distribution of bank account ownership, revealing an imbalance between individuals with and without bank accounts. To prepare the data, it was discovered that the data did not have any missing values, which implies that no imputation was necessary. The identifier column (uniqueid) was dropped since this is not a predictive variable. OneHotEncoder was used to encode categorical variables in a scikit-learn pipeline for country, education level, marital status, and job type, which ensures consistency between the training and test data. Numerical variables were transferred as is.

A Random Forest Classifier was chosen as the model of supervised learning because it is strong, able to deal with non-linear relationships, and has a high level of baseline. The data was divided into a training and a validation set, and the model performance was assessed in terms of accuracy.

Challenges Faced

Some difficulties were experienced in the project. First, coding of categorical variables through manual means introduced error because of non-observed categories in the test dataset. The solution to this problem involved the use of a pipeline that used OneHotEncoder to safely deal with unknown categories. The other problem was that of producing the submission file, where an error in the length was created by the discrepancy between the sample submission file and the test dataset. This was resolved by building out the submission using the unique identifiers of the test dataset. The difficulties have underscored the need to exercise correct preprocessing and meticulous alignment of datasets when carrying out actual machine learning tasks.

Key insights about financial inclusion

The model and the exploratory analysis established that the level of education, age, type of employment, and access to a mobile phone are important factors that may either lead to or not lead to the opening of a bank account. Educated ones and those who are under steady employment had better chances of being financially incorporated. Also, the availability of a cellphone became a significant indicator, which signals the increased role of digital and mobile-based financial services in Africa.

Lessons Learned

This project gave me a good practical experience in the application of supervised machine learning to a real-life issue. Some of the lessons learned were the role of preprocessing data, pipelines as the basis of reproducible systems, and the appropriate evaluation of models. Being part of Zindi also exposed me to joint workflows in data science and revealed to me the community-based problem-solving. In general, this challenge enhanced my perception of machine learning and its use to address socioeconomic development challenges.