SIMPLILEARN - MACHINE LEARNING COURSE

PROJECT 2

INCOME QUALIFICATION

WRITE UP

SUBMITTED BY

ADITHYA M. N.

B.TECH MECHATRONICS ENGINEERING

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

ABSTRACT

Latin America generally refers to territories in the Americas where the Spanish, Portuguese or French languages prevail: Mexico, most of Central and South America, and in the Caribbean, Cuba, the Dominican Republic, Haiti, and Puerto Rico. Many social programs have a hard time ensuring that the right people are given enough aid. It’s tricky when a program focuses on the poorest segment of the population. This segment of the population can’t provide the necessary income and expense records to prove that they qualify. In Latin America, a popular method called Proxy Means Test (PMT) uses an algorithm to verify income qualification. With PMT, agencies use a model that considers a family’s observable household attributes like the material of their walls and ceiling or the assets found in their homes to classify them and predict their level of need. The main aim of the project is to “Identify the level of income qualification needed for the families in Latin America.” The Inter-American Development Bank (IDB) believes that new methods beyond traditional econometrics, based on a dataset of Costa Rican household characteristics, might help improve PMT’s performance. The METHODOLOGY of the project has been mentioned below.

METHODOLOGY

To tackle this problem being faced, the following methodology has been followed. Two datasets have been provided namely ‘test.csv’ and ‘train.csv’. The various columns in the dataset provided were then studied and understood using the ‘Data Dictionary’ provided. The main data values were concluded to be from,

Id	A unique identifier for each row.
Target	The target is an ordinal variable indicating groups of income levels. ● 1 = extreme poverty ● 2 = moderate poverty ● 3 = vulnerable households ● 4 = non vulnerable households
idhogar	This is a unique identifier for each household. This can be used to create household-wide features, etc. All rows in a given household will have a matching value for this identifier.
parentesco1	indicates if this person is the head of the household

- We first import all the necessary libraries to ensure optimal functioning of the classifier. This is the first step. Then we load the dataset into the program using pandas. We can now perform various operations to perform the data.
- The first task was to identify the output variable. This is done by looking at the various columns of the dataset and finding out the level of the household.
- The data was understood by looking at the various types of data provided with us with the dataset.
- The biases of the dataset were identified by looking at the values counts of the targets and the difference in magnitude of the various values present in it was observed.
- The next step is to preprocess the data , finding all the columns which have null values in them. The Goal is to remove them because sklearn does not accept null values to train any dataset. Another methodology pursued was to keep the data which has more than 50% data as values, I have removed the data which has more than 50% null values. As they wont have a significant impact on the data. Then the remaining data with null values less than 50% were imputed, This converts the null values as zero as previously mentioned about the learning model of sklearn.
- Check whether all members of the house have the same poverty level. This was done by grouping the data with 'idhogar' and finding the number of unique values.
- Check if there is a house without a family head. This was done by grouping the data with 'idhogar' where the sum of the columns 'parentsco1' is not of a significant value.
- Set the poverty level of the members and the head of the house within a family. This was done by grouping the data with 'idhogar' where the mean of the target was used to identify the column and perform the operation.
- Predict the accuracy using a random forest classifier. The classifier was first imported from the sklearn ensemble library. A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. This Classifier is highly recommended due to its high accuracy and ability to not overfit the data. The data was

split into train and test data with 80% to 20% value ratios. The data was fit and the classifier was then trained.

- The accuracy for various predictions was tested and the random forest classifier was found to classify the data with a 93% accuracy. The accuracy of the Random Forest Classifier was then cross validated using the ‘KFold cross validation’. The performance of the classifier here was found to have a mean score 94.7% accuracy.

CONCLUSION

The project was successfully able to “Identify the level of Income Qualification required by the people of Latin America”. This was done with the help of a “Random Tree Classifier” with an accuracy score of 93% for the classifier. This classifier was also cross validated with KFold Cross-Validation.