SocialCops ML Land Classification Challenge

Royal Bhati 7-11-2018

First three things:

1. I looked at the first few rows of the dataset to have a some understanding about what kind of data I was working on using pandas library.
2. After looking at the data I summarised the data to check statistical measures of the data.
3. After summarizing I drew some graphs and plots to uncover the patterns and relationship in the data

All the three things are documented in the jupyter notebook and the html file.

Preprocessing Involved:

1. Data was highly imbalanced so I resampled the data so that It should not be biased towards the frequent occurring class.
2. After resampling I shuffled the data because the data was sorted based on the target and the train and validation split may not generalize the overall data
3. Normalizing or scaling the variables wasn't required because tree-based models can work on unnormalized data pretty well as in scaling and normalizations we are just dividing/multiplying or adding with a constant and that doesn't provide any relevance.

Models:

I tried tree based models because they give better results(mostly) as compared to non-tree based.

1. Random Forest
2. XG Boost

I could have tried an ANN model but the data was simple and tree-based models were giving good results.

1. Random Forest: A Random Forest consists of a collection or ensemble of simple tree predictors, each capable of producing a response when presented with a set of predictor values. For classification problems, this response takes the form of a class membership, which associates, or classifies, a set of independent predictor values with one of the categories present in the dependent variable.

2. Xgboost :XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM)

Error Measurement :
Since there was no evaluation metric given so I used the ***multi class Log loss*** metric to train the model in Xgboost and **Accuarcy** in Random Forest to evaluate the results.

What could have been done?

Since the data dictionary was not given I could have gone much deeper to get insights about the columns and what they describe.

Additional Comments:

Since there were no column names and column description feature engineering couldn't be done so I had to start with the simple approaches.

After working with dataset there was a certain impression that some of the preprocessing was done already.