Please choose two projects from these projects:

Project 1: Domain – Social Media

FOCUS – PREDICT NO. OF SHARES OF AN ARTICLE

BUSINESS CHALLENGE/REQUIREMENT

Mashable (www.mashable.com) -- is a global, multi-platform media and entertainment company. Powered by its own proprietary technology, Mashable is the go-to source for tech, digital culture and entertainment content for its dedicated and influential audience around the globe.

Just like any other media company its success depends on the popularity of articles. And one of the key metrics to measure popularity is no. of shares done on article.

Over period of few years Mashable has collected data on around 40,000 articles. You as ML expert have to do analysis and modeling to predict number of shares of an article given the input parameters.

BUSINESS BENEFITS

Mashable's entire business is dependent on popularity of articles. With accurate prediction of shares, company can choose which articles to publish hence driving higher user engagement and profits. Rough estimate is 1% increase in engagement time (minutes) increases profit by up to 5%.

Project 2: Domain - Agriculture

New DG Food Agro are a multinational exporter of various grains from India since nearly 130 years. But their main product of exporting since early 1980s has been Wheat. They export wheat to countries like America, Afghanistan, Australia etc.

They started seeing varying exports of sales year on year for various countries. The reason that was theorized by them had a lot of natural causes like floods, country growth, population explosion etc. Now they need to decide which countries fall in the same range of export and which don't. They also need to know which countries export is low and can be improved and which countries are performing very well across the years.

The data provided right now is across 18 years. What they need is a repeatable solution which won't get affected no matter how much data is added across time

and that they should be able to explain the data across years in less number of variables.
Objective: Our objective is to cluster the countries based on various sales data provided to us across years. We have to apply an unsupervised learning technique like K means or Hierarchical clustering so as to get the final solution. But before that we have to bring the exports (in tons) of all countries down to same scale across years. Plus, as this solution needs to be repeatable we will have to do PCA so as to get the principal components which explain max variance.

Implementation:

1) Read the data file and check for any missing values
2) Change the headers to country and year accordingly.
3) Cleanse the data if required and remove null or blank values
4) After the EDA part is done, try to think which algorithm should be applied here.
5) As we need to make this across years we need to apply PCA first.
6) Apply PCA on the dataset and find the number of principal components which explain nearly all the variance.
7) Plot elbow chart or scree plot to find out optimal number of clusters.
8) Then try to apply K means, Hierarchical clustering and showcase the results.
9) You can either choose to group the countries based on years of data or using the principal components.
10) Then see which countries are consistent and which are largest importers of the good based on scale and position of cluster.

Project 3: Domain - Agriculture

Maple Leaves Ltd is a start-up company which makes herbs from different types of plants and its leaves. Currently the system they use to classify the trees which they import in a batch is quite manual. A labourer from his experience decides the leaf type and subtype of plant family. They have asked us to automate this process and remove any manual intervention from this process.

Objective: To classify the plant leaves by various classifiers from different metrics of the leaves and to choose the best classifier for future reference.

Implementation:

1) Import the train and test csv.
2) Import the required classification libraries along with pandas, numpy, seaborn etc
3) Then import the classifiers from them (Randomforest, SVM, NaiveBayes, DecisionTrees)

4) After this create a function to encode the labels of the strings given in the dataset

5) You can do the above step using label encoder. With this you are creating some labels from train set as test set. The test set we imported is for testing the best classifier accuracy once we choose it

6) Then extract the values from train set by stratifying them and dividing it into 80:20 ratio

7) Now your X train, X test, Y train, Y test are ready.

8) We currently don’t know which is the best classifier on the dataset. So, we apply all 4 of them.

9) Create the classifiers class and initialize all the respective classifiers

10) Then run the X train & X test datasets through classifiers calculating the log loss and accuracy of the result

11) Choose the classifier which has the best accuracy

12) Then try to predict the result on the import test.csv dataset