SD-701 Bigdatamining

Report

The goal of this document is to sum up the principal actions done in order to study the dataset Given. It is composed of 56 column desxcribing caracteristics of trees in different forest. The goal there is to predict the cover type given the infomation of the tree.

1 Dataset Exploration

The first things done was to see the repartition of the data and find out that some classes are way more represented than others. Then I plotted the different variables and their potential correlation. I saw that some of may be transformed in order to improve our models later.

After a first exploration of the data, i tried the first model

2 First model

The first model tested was the model given as an example. The simple logistic regression gave a 0.7 score on Kaggle. I tried to improve it by changing the parametters, but it appears quite fast that itw as not the best way to classify our data.

That’s why i tried to apply other models.

3 Test of different models

I first improved the model i tested by using cross validation, to different parameters on my model in order to find the best result it could give to me. Logistic regression was worse than random forest which was himself worse than Decision Tree after differents tries.

4 Improve the best model

After deciding which model i would try to improve more (Decision Tree), i found out the best parameters to use to get the best score on kaggle.

[Model 1] Decision Tree Kaggle score = 0.90633

- maxDepth = 30
- maxBins = 200
- impurity = entropy
- minInstancesPerNode = 3

After trying this model, i tried to find out if I could improve my result using the library Sklearn. Indeed, this library is well optimised and provide a lot of model not reachable with Mllib

5 Sklearn model

After converting the data to pandas dataframe, i tried my first sklearn models. I found out the powerfullness of the library : even a basic random forest model was way better than my best Decision tree model. So i decided to find the best model and found the ExtraTreesClassifier.

I tunned some parameters, but I was quite fast limited by the cluster provided by databriks. I could not increase the number of estimator to more than 210 (approximatly) without exeeding the limit of my cluster **[Model 2]** . That's why i decided to execute it locally on a Jupyter notebook.

When I improved the parameters of the extraTrees locally I reached a precision of 0.945. To progress beyond that I decedid to work on the data again :

- I droped the ID column since it does not give any important information to the model
- I did feature engineering adding some column (some because the seems logical such as the distance, some because i knew that the classifier was spliting data on a threshold, so i needed to make them more straight). It did iproved my score noticiably.
- I normalised the data (scaling them, centering and so on)
- I selected the features the most important for the model regarding to the pvalue
- Limitting the number of features in the model was also a good improvement
- Use AdaBoostClassifier to finally improve a last time

The final model, which gave me my best result :

[Model 3] ExtraTrees : Kaggle score = 0.95967

- n_estimators = 270 (if we use too much we risk overfitting)
- criterion= 'entropy'
- n_jobs = -1 (to use all the CPU)
- warm_start = True
- max_features = 19

As a conclusion, the score can for sure still be improved. Finding other new features to add would probably give me best results such as optimising more my model.

File given

You will find in this git repo 2 files containing the code of those model as well as the results given by my three models :

- « SDI_701_mllib_Valentin_Larrieu.ipynb» which contains the code of **model 1** and **model 2**
- « SDI_701_sklearn_Valentin_Larrieu.ipynb» which contain the code of **model 3**
- « SDI_701_model1_Results _Valentin_Larrieu.csv» which contains the results given by **model 1**
- « SDI_701_model3_Results _Valentin_Larrieu.csv » which contains the results given by **model 3**