ERHVERVSAKADEMI AARHUS
BUSINESS ACADEMY AARHUS

Exercises Week 7

ITA Machine Learning.

The exercises this week are a bit more “open-ended” than you have seen before. But use the techniques that you have been introduced to, so far, in this course,

Work in groups of 2-3.

Exercise.

A bank deals in home loans. They give loans across urban, semi-urban and rural areas.

Customers must send information about Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. To automate this process, the bank then wants a program that can identify customer segments, those who are eligible for loans, and those that are not.

The bank provides a training dataset with loans given, as well as a test dataset, where your program should be able to decide (whether to grant a loan or not) with the highest possible accuracy.

Information given:

Dataset Description:

Variable	Description
Loan_ID	Unique Loan ID
Gender	Male/ Female
Married	Applicant married (Y/N)
Dependents	Number of dependents
Education	Applicant Education (Graduate/ Under Graduate)
Self_Employed	Self employed (Y/N)
ApplicantIncome	Applicant income
CoapplicantIncome	Coapplicant income
LoanAmount	Loan amount in thousands
Loan_Amount_Term	Term of loan in months
Credit_History	credit history meets guidelines
Property_Area	Urban/ Semi Urban/ Rural
Loan_Status	Loan approved (Y/N)

Notice:

Loan Approval Status: About 2/3rd of applicants have been granted loan.

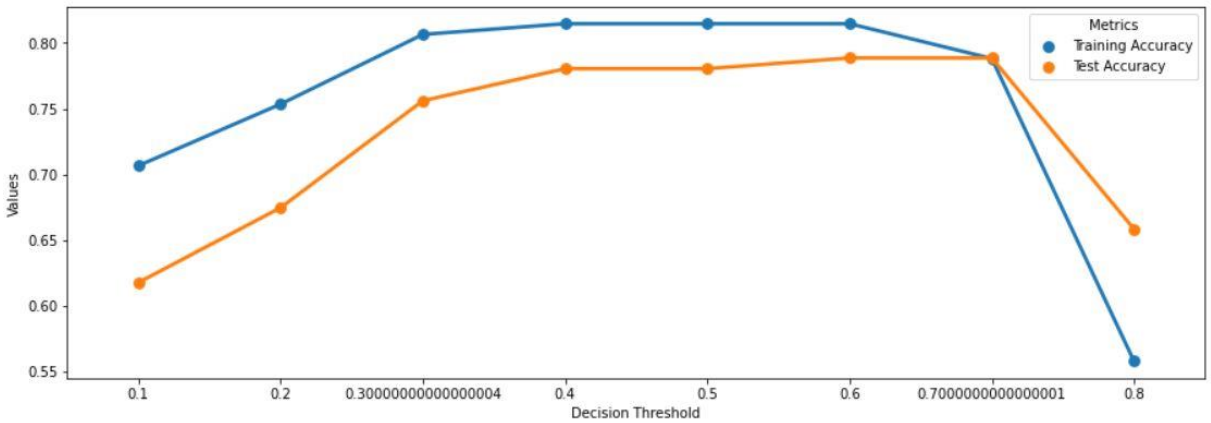
Education: About 5/6th of the population is Graduate and graduates have higher propotion of loan approval.

Employment: 5/6th of population is not self employed.

Property Area: More applicants from Semi-urban are likely to be granted loans.

Applicants with a credit history are far more likely to be accepted.

We then split the original dataset into a training set and a test set (with a 80/20 split). Using logistic regression, we measure our accuracy with various thresholds.



A threshold somewhere between 0.3 and 0.6 appears to be optimal.

a) Verify that all in the group understand the code ("Loan Approval Prediction.ipynb")
b) Next, make similar plots using sklearns
   - decision tree or random forest.
   - neural net.

So, can you get similar results with a random forest classifier? What about the neural net? Is it possible to improve the results from the code you were given (that used logistic regression)? What are the best values for the random forest parameters (number of trees, tree depth), and the neural net parameters (number of hidden layers, number of neurons etc.)?