CS 412 – Homework 2

Fatih Tasyaran
19321

August 6, 2019

1 Introduction

For this homework, I used dataset bank-additional-full.csv which provided in homework. Labeled all data points in proper format in order to use with Decision Tree Classifier, and tried depths from 1 *to* 50 and $full\_depth$ on all versions I came up with. To observe depth's effect, and feature's importance's effect on accuracy of model.

2 Input Versions

As a feature of the dataset, I buit different Decision Trees with and without existence of the feature "*duration*". As the bank-additional-names file mentioned, this feature highly affects the output target. For both test my implementation's correctness and to observe feature's and value's effect on accuracy of model, I tailored the data set in various ways. At the end, I had 5 different subset of the input which are:

2.1 Highly Tailored Subset

In this subset, I exclude "*duration*" column from the data. I also excluded every row which includes *unknown* and *nonexistent* in their columns. My aim was to reduce noise in the data as much as possible. But after this exclusion, $\tilde{4}0000$ rows of data reduced to $\tilde{4}000$ rows.

2.2 Highly Tailored Subset with *Duration*

I prepared this subset to observe "*duration*"'s effect on a hopefully less noisy and smaller subset. This subset is same *Highly Tailored Subset*, but with *duration* feature.

2.3 All Data Points

In this subset, I didn't exclude any of the data points from the data, my aim was to get a baseline for a noisy model to compare accuracy with other inputs. Also this subset is useful to see importance of features other than *duration* in a big and noisy data.

2.4 All Data Points Except *Duration*

In this subset, my aim was to get a baseline for all data provided, without the noise of *duration*. This subset is for comparing to my *logical cut*.

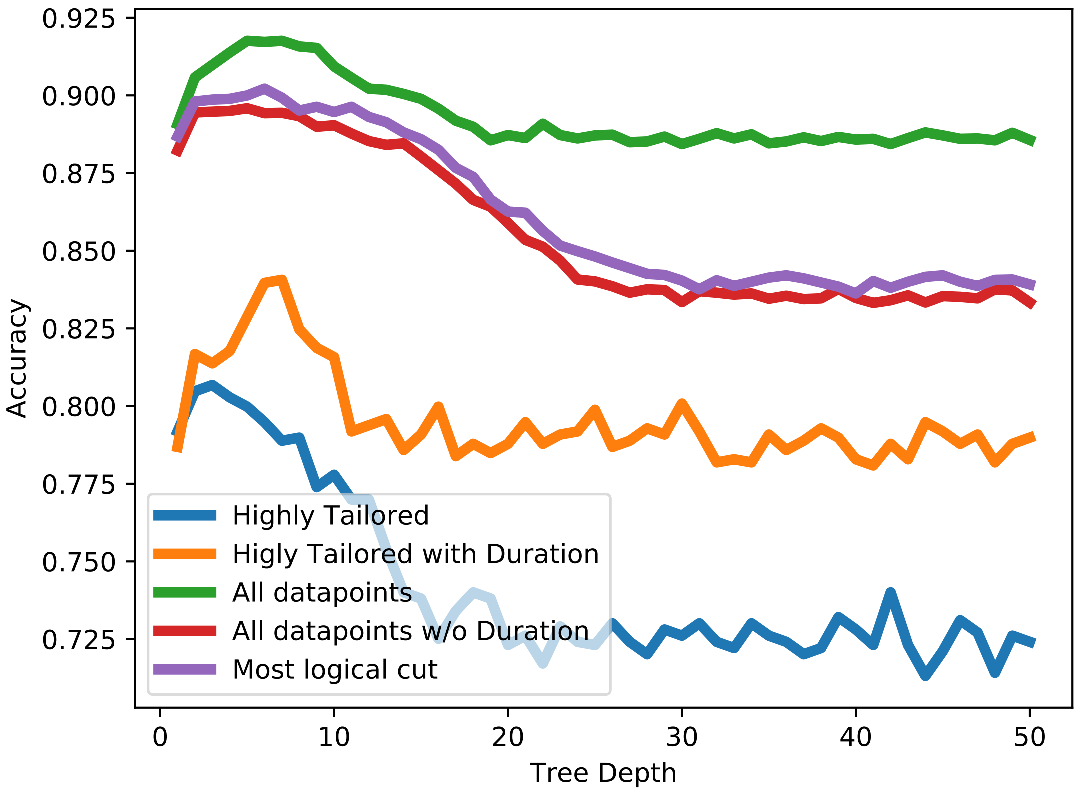
2.5 Most Logical Cut

In this subset, I aimed to get most general and logical input. I excluded *duration* since it have tremendous effect on target value, I also excluded day_of_week since it seems to have no logical effect on a deposit application. But I keep *month* since financial applications generally affected by period of the year. I also kept *unknown* and *nonexistent* values to maintain volume of the data.

3 Implementation

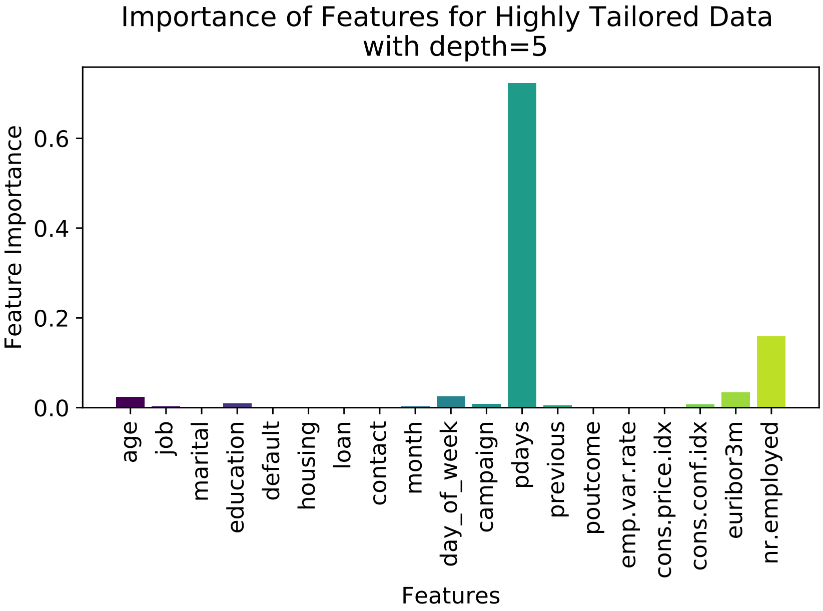
I used *DecisionTreeClassifier* from *sklearn* to implement classifier. *LabelEncoding* from *sklearn.preprocessing* in order to convert data to integer based categorical, instead of string based categorical which *sklearn.tree* dind't accept as parameter.
I used *train_test_split* from *sklearn* to implement cross-validation. I used 0.2 test size and split data only once for every run from depths 1 *to* 50. I fit and run the classifier from depths 1 *to* 50 and *max_depth* for each subset of data set.

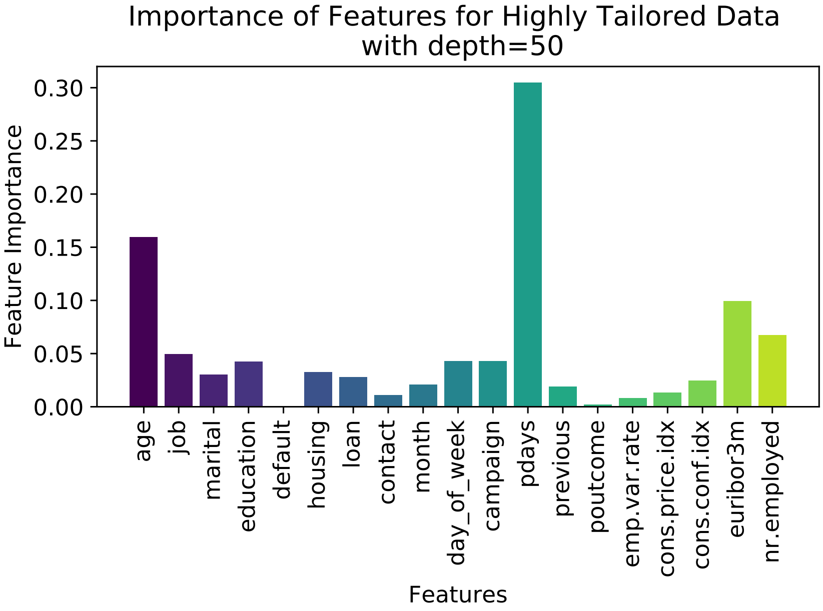
4 Results & Discussion

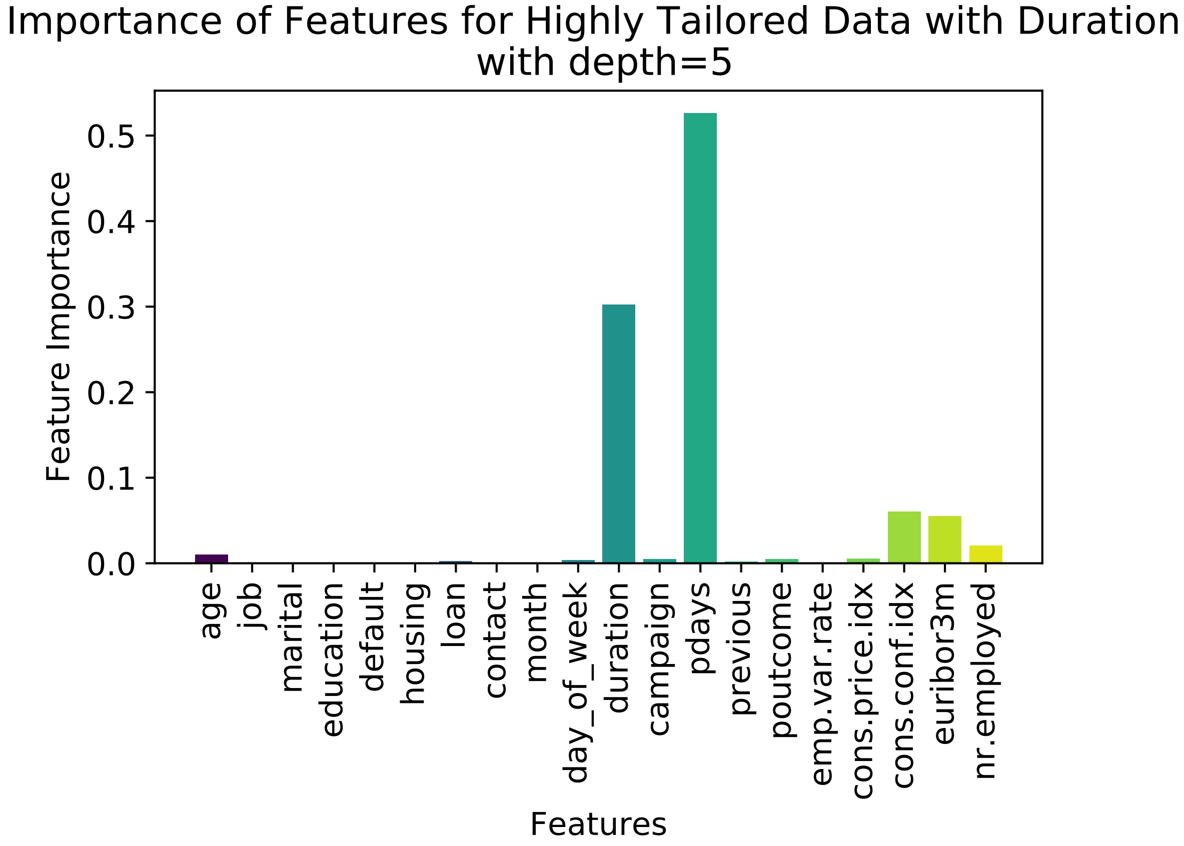


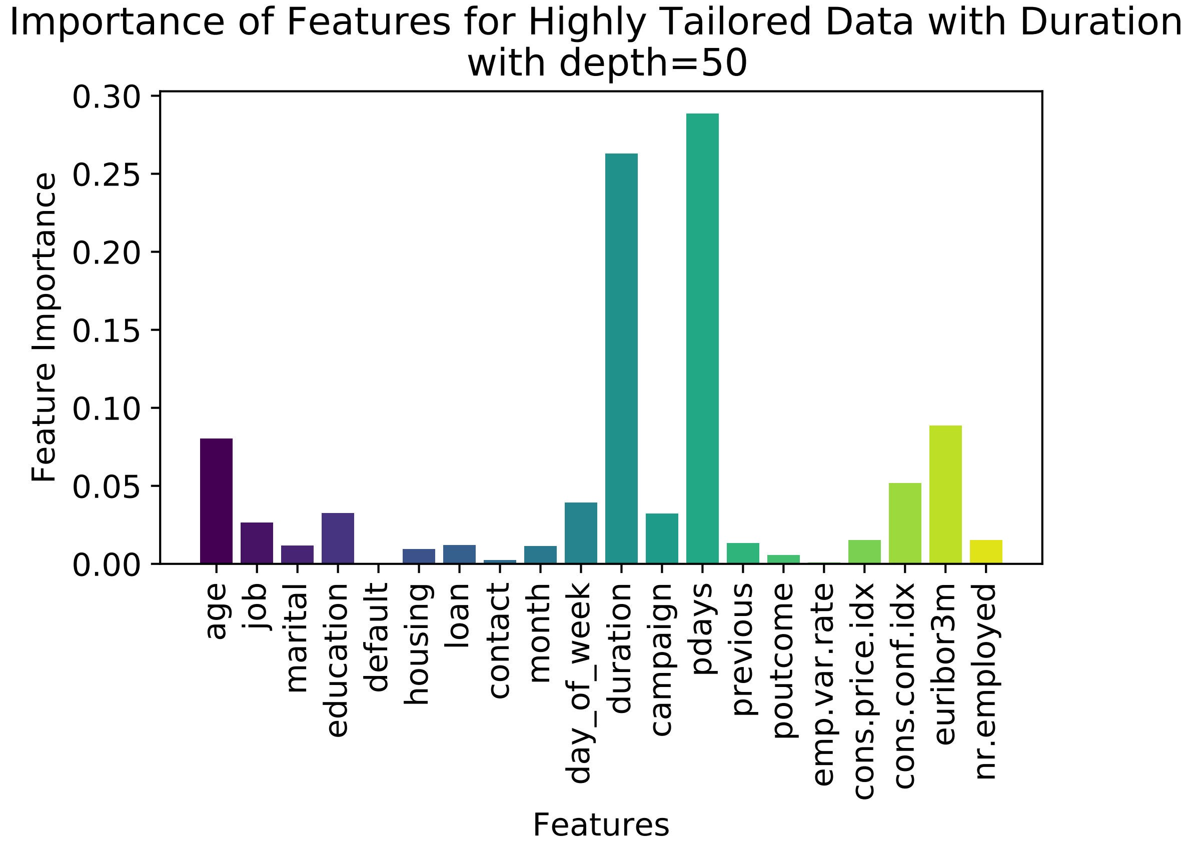
My observations on accuracy of models trained with different subsets of same data are:

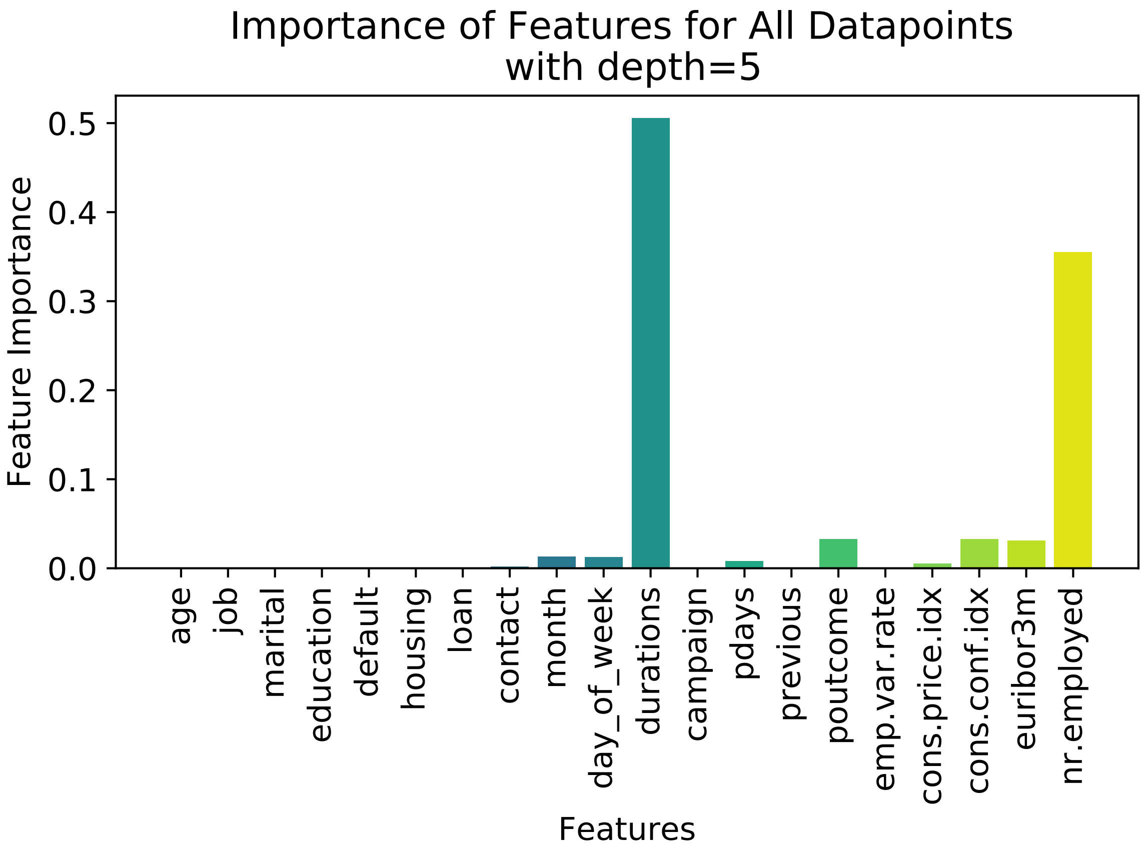
- Highly tailored subset performs worse among all subsets. I think this is due to low volume of remaining data after tailoring.
- Subsets with *duration* perform better than their counterparts, this both proves correctness of implementation and over fitting in subsets including *duration*.
- Most logical cut performs better than all data points w/o duration, which is actual baseline for model correctness provided with data set. Proves that, there are unnecessary and noise causing data points in the dataset such as day_of_the_week. Discarding them improves correctness of model. all models, trained with different subsets, accuracy decreases as depth of tree increases, in order to investigate this behaviour, I looked at feature importances of all models with $depth = 5$ and $depth = 50$

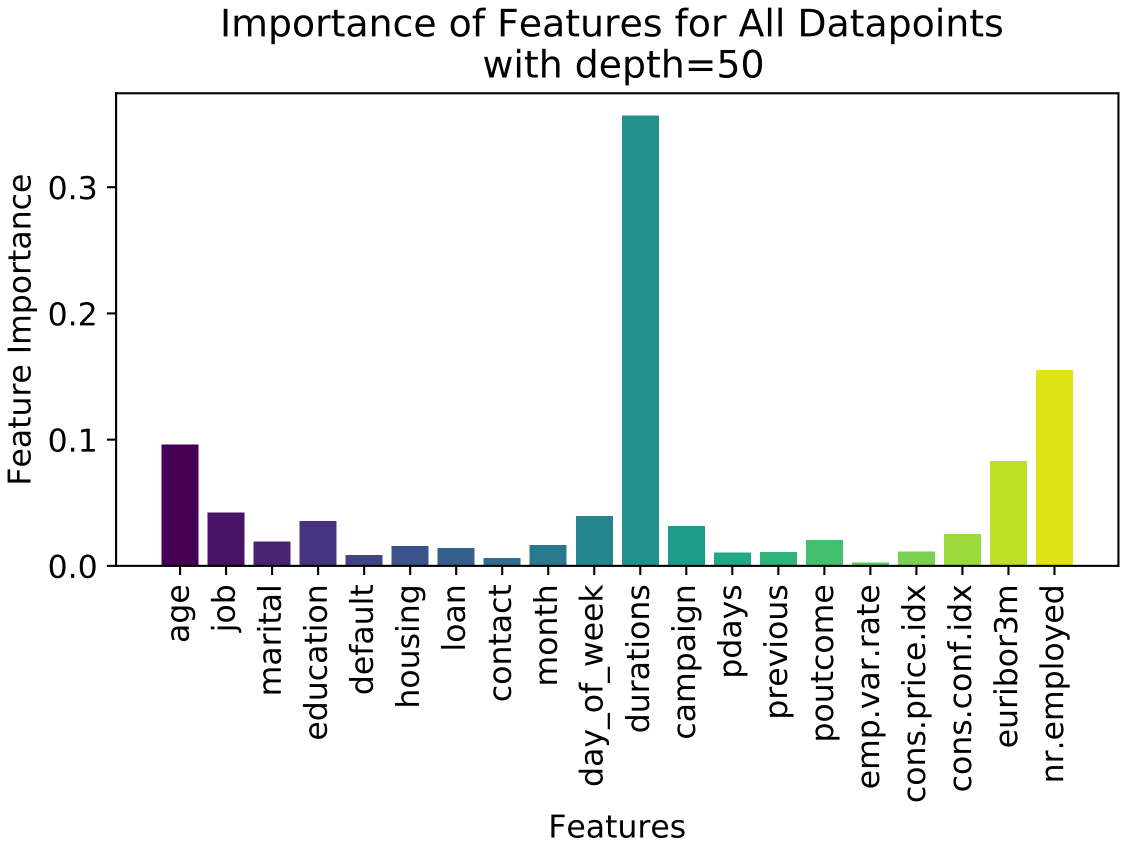


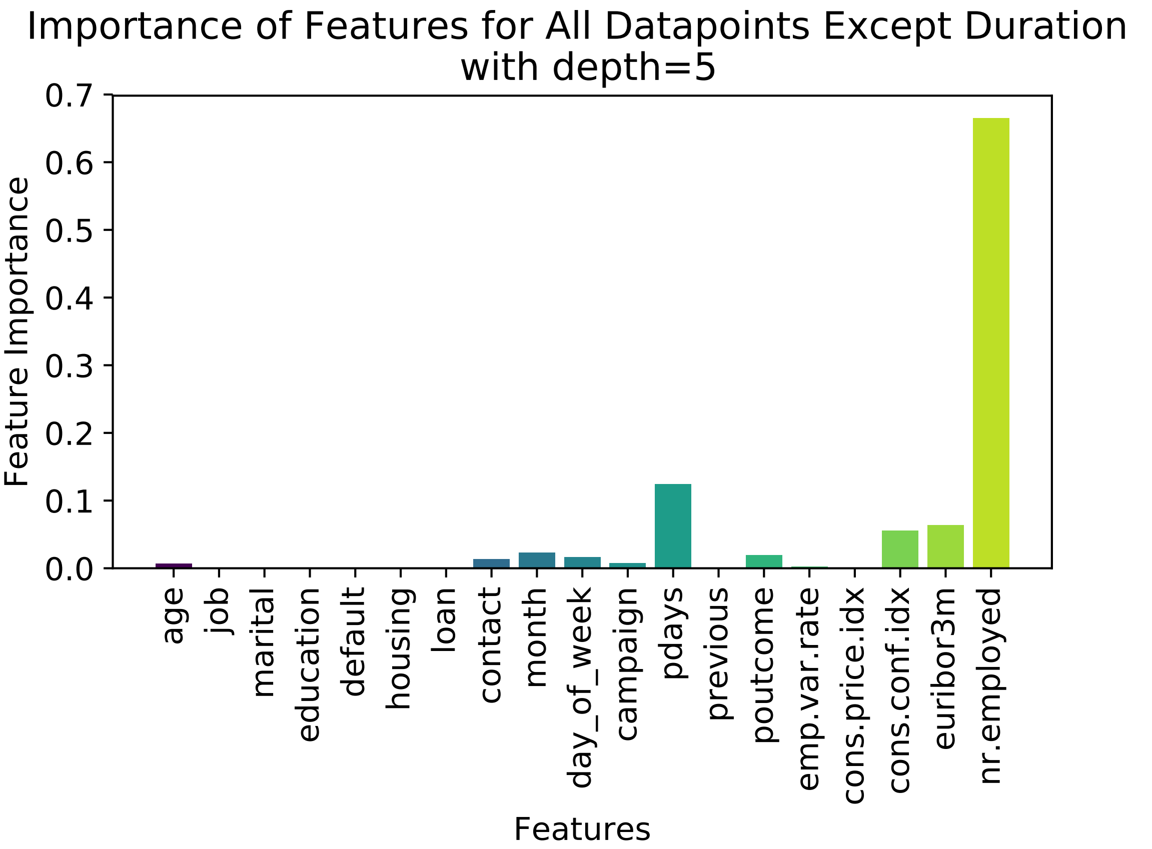


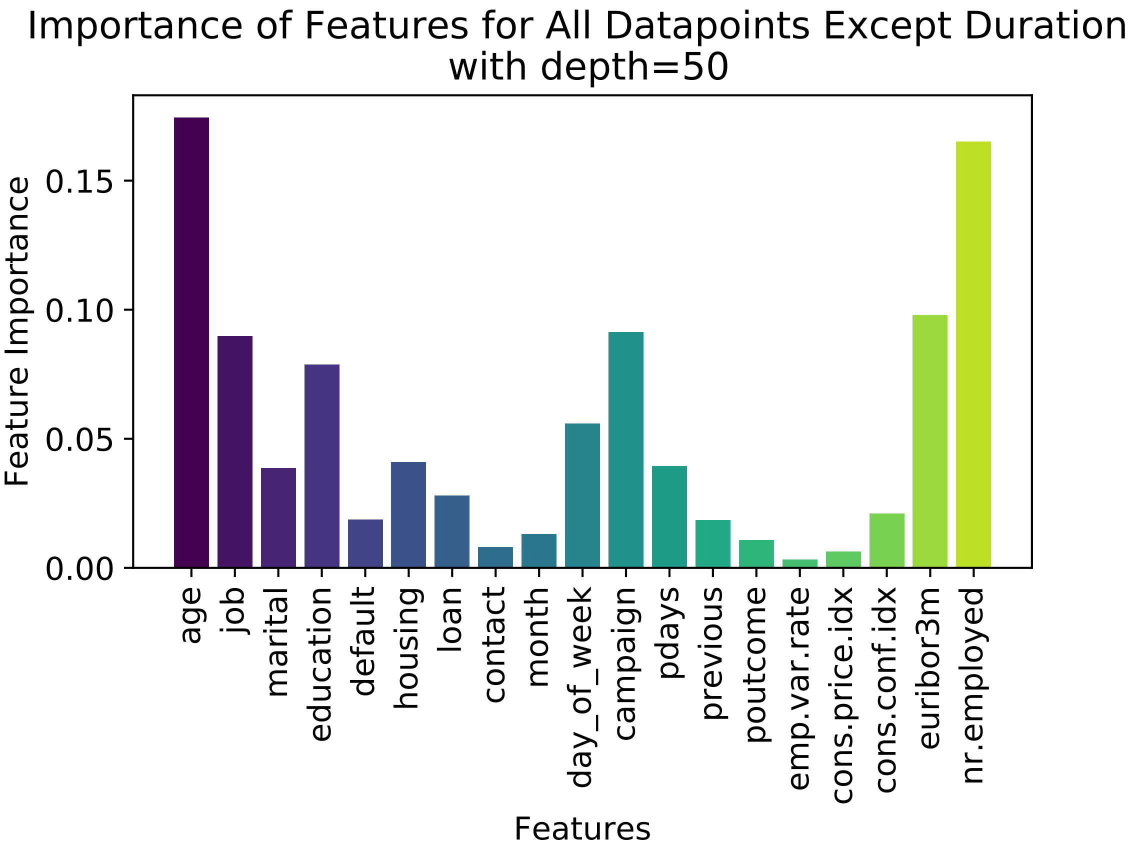

Importance of Features for Highly Tailored Data with Duration with depth=5
Feature Importance
0.0
0.1
0.2
0.3
0.4
0.5
age
job
marital
education
default
housing
loan
contact
month
day_of_week
duration
campaign
pdays
previous
poutcome
emp.var.rate
cons.price.idx
cons.conf.idx
euribor3m
nr.employed
Features

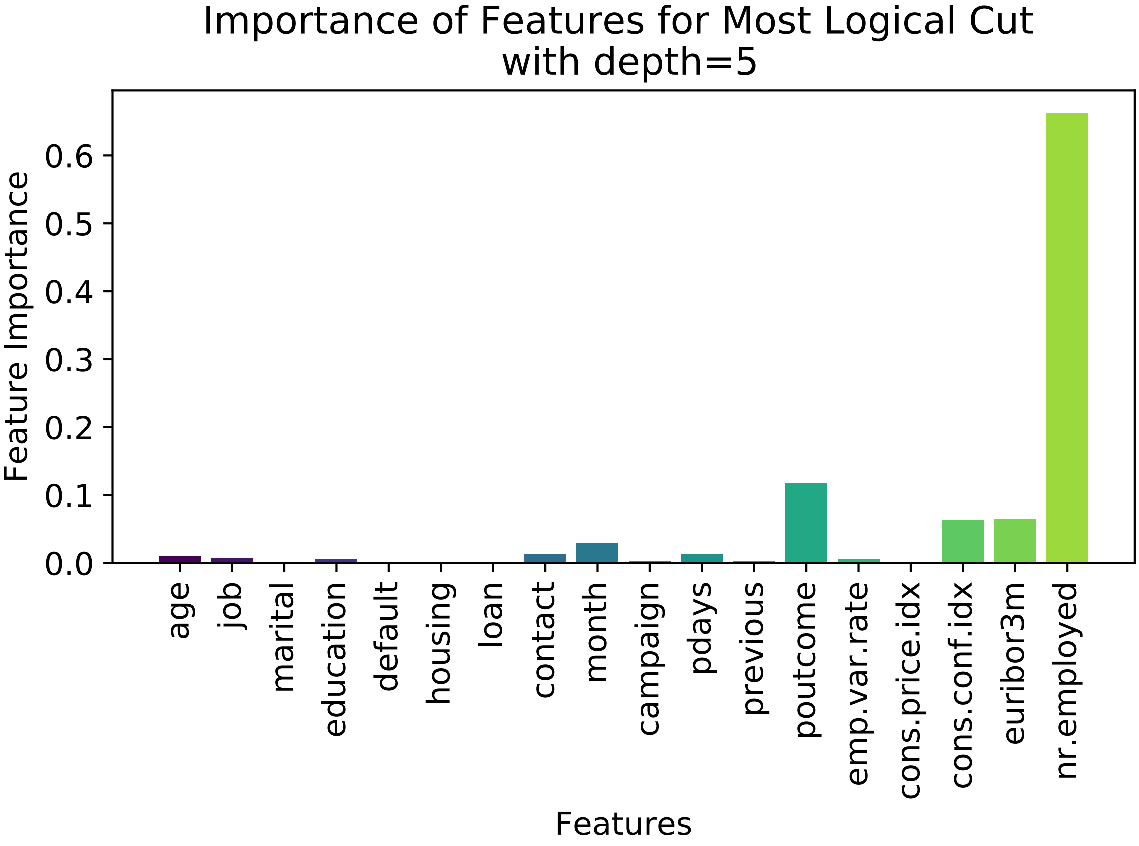

Importance of Features for Highly Tailored Data with Duration with depth=50
Feature Importance
0.00
0.05
0.10
0.15
0.20
0.25
0.30
age
job
marital
education
default
housing
loan
contact
month
day_of_week
duration
campaign
pdays
previous
poutcome
emp.var.rate
cons.price.idx
cons.conf.idx
euribor3m
nr.employed
Features

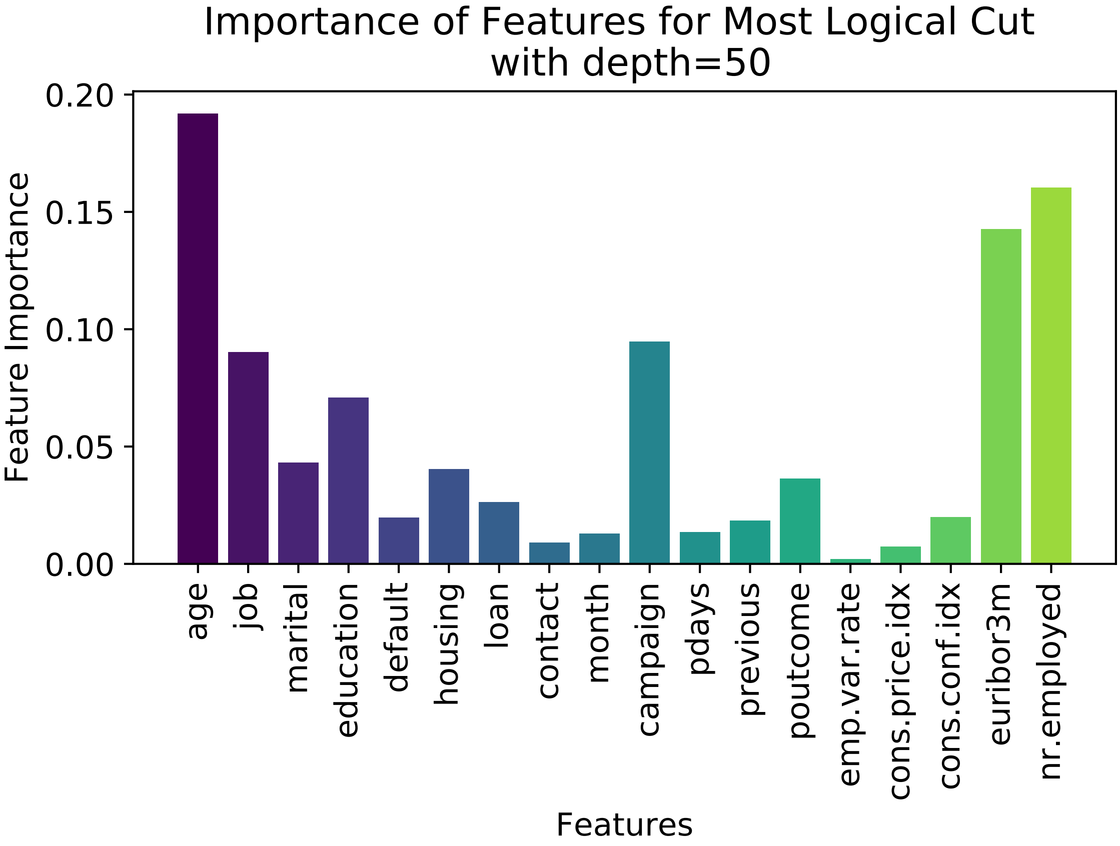

Importance of Features for All Datapoints
with depth=5
Feature Importance
0.0
0.1
0.2
0.3
0.4
0.5
age
job
marital
education
default
housing
loan
contact
month
day_of_week
durations
campaign
pdays
previous
poutcome
emp.var.rate
cons.price.idx
cons.conf.idx
euribor3m
nr.employed
Features


Importance of Features for All Datapoints
with depth=50
Feature Importance
0.0
0.1
0.2
0.3
age
job
marital
education
default
housing
loan
contact
month
day_of_week
durations
campaign
pdays
previous
poutcome
emp.var.rate
cons.price.idx
cons.conf.idx
euribor3m
nr.employed
Features


Importance of Features for All Datapoints Except Duration with depth=5
Feature Importance
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
age
job
marital
education
default
housing
loan
contact
month
day_of_week
campaign
pdays
previous
poutcome
emp.var.rate
cons.price.idx
cons.conf.idx
euribor3m
nr.employed
Features


Importance of Features for All Datapoints Except Duration with depth=50
Feature Importance
0.00
0.05
0.10
0.15
age
job
marital
education
default
housing
loan
contact
month
day_of_week
campaign
pdays
previous
poutcome
emp.var.rate
cons.price.idx
cons.conf.idx
euribor3m
nr.employed
Features


Importance of Features for Most Logical Cut
with depth=5
Feature Importance
0.0
0.1
0.2
0.3
0.4
0.5
0.6
age
job
marital
education
default
housing
loan
contact
month
campaign
pdays
previous
poutcome
emp.var.rate
cons.price.idx
cons.conf.idx
euribor3m
nr.employed
Features


Importance of Features for Most Logical Cut
with depth=50
Feature Importance
0.00
0.05
0.10
0.15
0.20
age
job
marital
education
default
housing
loan
contact
month
campaign
pdays
previous
poutcome
emp.var.rate
cons.price.idx
cons.conf.idx
euribor3m
nr.employed
Features

- It could be seen by importancce of features that, increasing depth of the tree is causing overfitting to the data as it increases importance of all features, makes model more complicated than data.
- The effect of duration on target value is also highly visible in feature importances. It's a big amount that make model more consistent over increasing depth. Due to preventing other features become more important.

For further investigation. I also draw decision trees of each model at *depth* = 5 almost all of them perform their best.

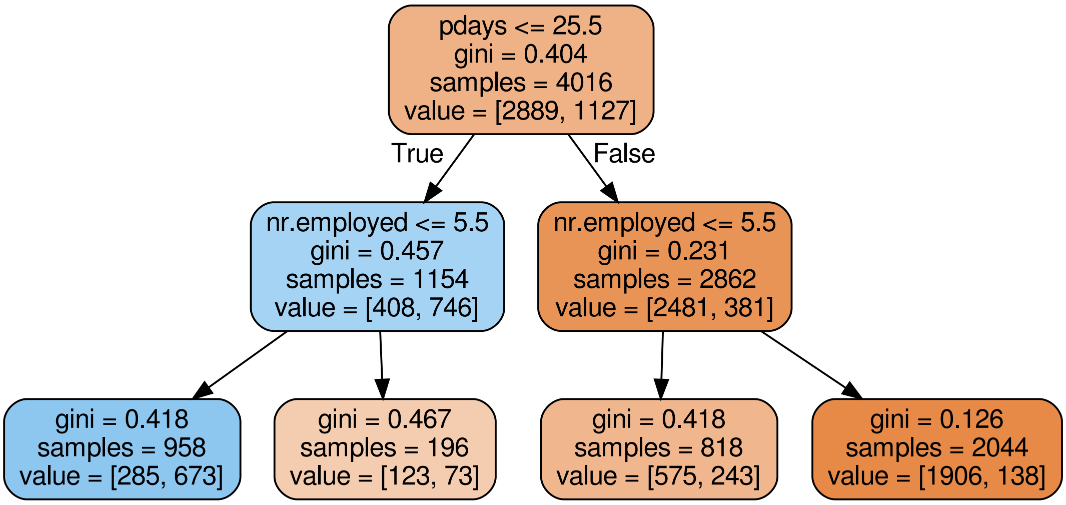


Figure 1: Tree of Highly Tailored Subset at *depth* = 2

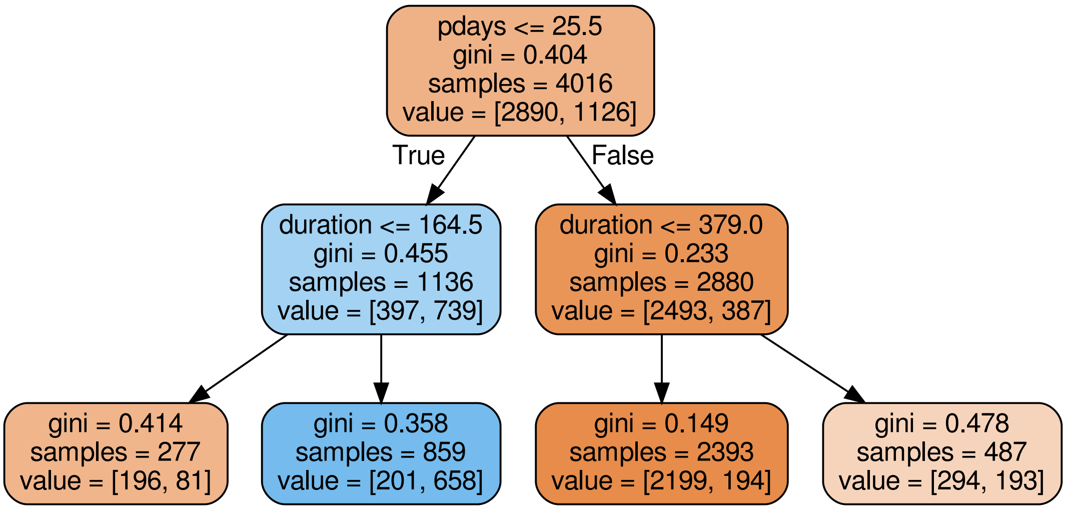


Figure 2: Tree of Highly Tailored Subset with *duration* at *depth* = 2

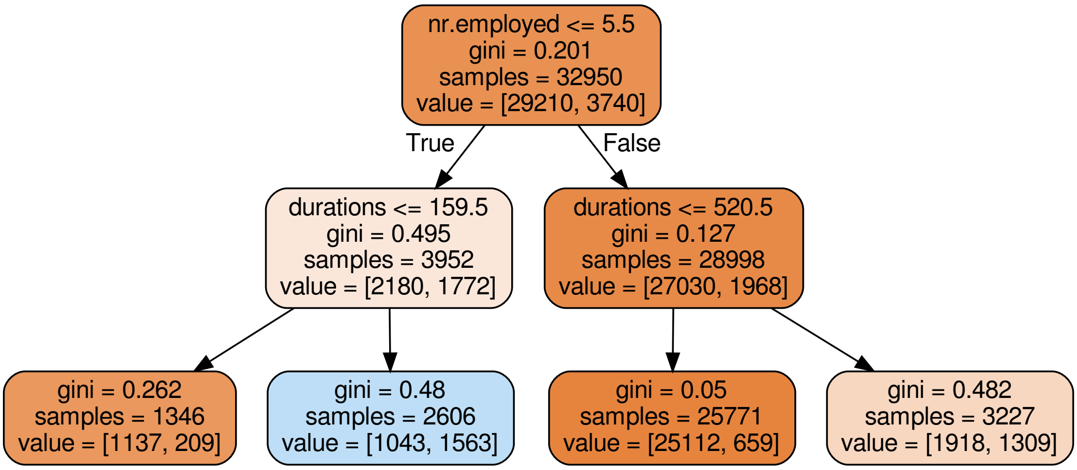


Figure 3: Tree of All Data Points at *depth* = 2

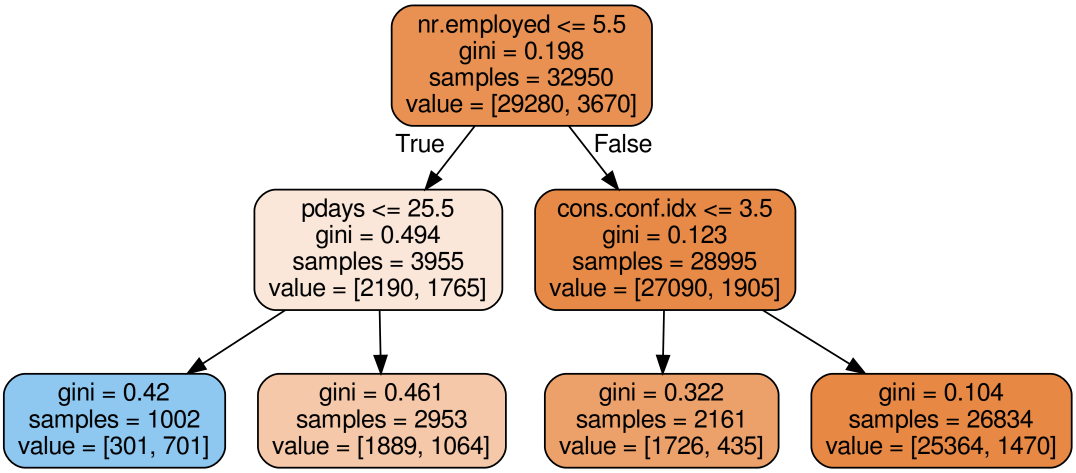


Figure 4: Tree of All Data Points except *duration* at *depth* = 2

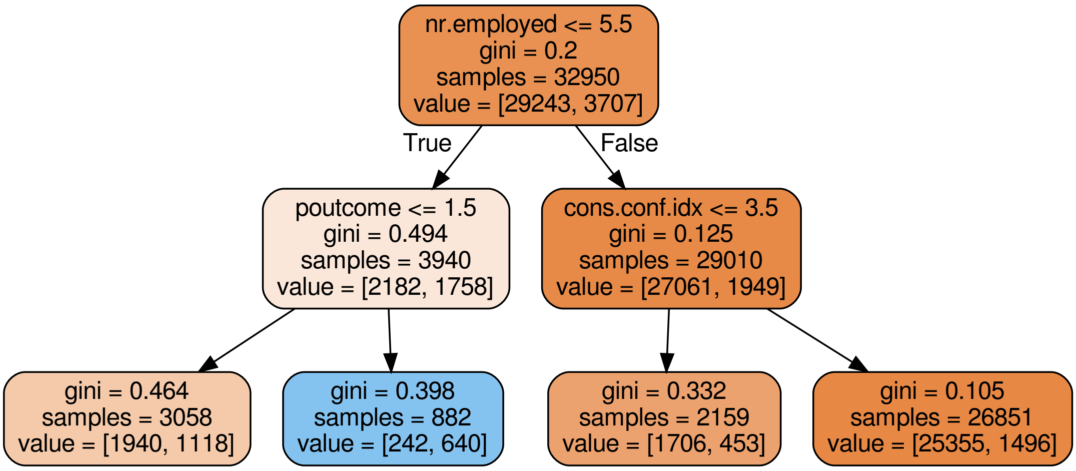


Figure 5: Tree of Most Logical Cut at $depth = 2$

- The effect of *duration* again clearly seen in decision trees and explains the accuracy rate difference with models that don't include *duration*.
- Duration rapidly decreases gini as could bee seen in *Figure 3* which results in high accuracy rates.
- In subsets cleared from unknown data points, *pdays* gains important, make thought all other features are known if *pdays* is known.
- In subsets *duration* don't exist, *nr.employed* gain importance, as accordance with finance logic.

5 Notes

- I commented tree drawing lines in code, so it don't create lots of file in directory.
- I submitted bank-additional-full.csv with code itself, code is ready to run.
- Implementation is on Python3
- It will pop-up all of graphics included here, they come one by one after skipping.
- Thanks!