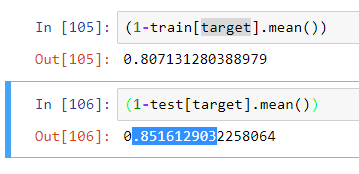
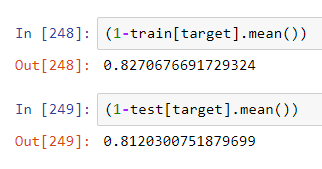
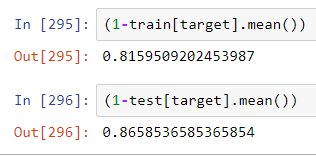
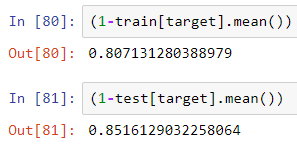
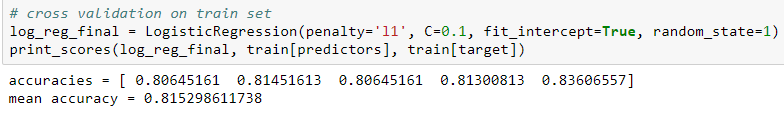
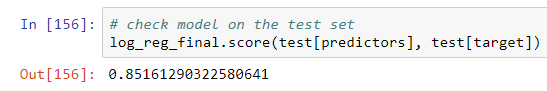
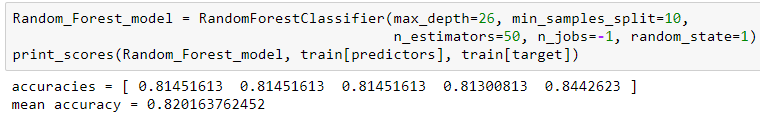
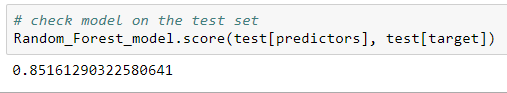
Notes:

* Dateset 0 – has both key1 and key2. It has some values and response as well
  + Response – 1 or 0
  + Keys and values - integer
* Dataset 1 – has only key2 and a lot of values for it
  + Key2 – integer
  + Values – float or null (nan / na). Majority are NaN
* Dataset 2 – hase only key1 and a lot of values for it
  + Key1 – integer
  + Values – mainly text(string). Some floats and integer. Some columns are (mainly) NaN
* drop\_NA\_only\_columns\_and\_rows – assumes no duplicates in column names
* 
* With col\_thresh (0.60) [42 cols], and row\_thresh (0.05) -> Random Forest gives 0.855 on test. LogReg 0.848
* With col\_thresh (0.70) [31 cols], and row\_thresh (0.05) -> Random Forest gives 0.8516 on test. LogReg 0.845
* With col\_thresh (0.73) [4 cols], and row\_thresh (0.05) -> Random Forest gives 0.8516 on test. LogReg 0.8516
* 
* With col\_thresh (0.60) [40 cols, 665 rows], and row\_thresh (0.30) -> Random Forest gives 0.8346 on test. LogReg 0.827
* 
* With col\_thresh (0.60) [44cols, 408 rows], and row\_thresh (0.70) -> Random Forest gives 0.878 on test. LogReg 0.8414
* Base case
  + col\_thresh=0.60, row\_thresh=0.05
  + majority class prediction:
  + 
  + Log\_reg\_l1:
  + 
  + 
  + Random Forest:
  + 
  + 

TODO:

* Task 2 - Determining key factors:
  + Add B0 coefficient?
  + v179 – 94 dummies created – drop that variable? Find a generic way
    - before creating dummies drop string columns that have to0 many unique categories
  + add graph showing the most important factors
  + text columns – replace ‘-‘, ‘+’, ‘ ’, … with ‘\_’
  + strings -> toLowerCase
  + drop columns with too little variation? (numeric) – little variation still can be useful
  + drop string columns with too many classes – check?
  + Remove FOR in Pandas to Orange conversion?
  + Perform correlation across factors. Plot them. Drop similar factors.
  + Using polynomial features?
  + Remove outliers
  + Show number of NAs
* Task 4 – improvements
  + Better handling of NAs

Questions:

1. Should we do inner join of task.app.data with task.cb.1 and task.cb.1? Or outer join?
2. When joining should we use key1+key2 as a unique key? Or we just separately add all values for key1 and then for key2?
   1. We should add separately all values for key1 and key2 because dataset 1 has only key2 and dataset 2 has only key1
3. Response variable is only in dataset0. So, for factor importance analysis and prediction we should use only inner join?