Data Understanding and Data Preparation Exercise

Courtesy of Intel Advanced Analytics

In this exercise you will explore data that is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be (or not) subscribed.

Data source: [Moro et al., 2011] S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology.

Good luck!

Q1

Familiarizing yourself with the data

1) Download the 'bank-full_2016.csv' database to a local directory, and load it into a DataFrame object: http://slavanov.com/teaching/ds1617b/bank-full_2016.csv (http://slavanov.com/teaching/ds1617b/bank-full_2016.csv)

In [ ]:

```
```

2) Explore your data a bit:

- What are the dimensions of the table?
- What are the different attributes in it?
- Return a frequency table of the "outcome" category in the table - i.e return all the possible values in that category, and how many times each value is present in the table.
- Display the first 5 rows of the table.

In [ ]:

```
```

3) What is the data type of each attribute in the table?

In [ ]:

```
```

Manipulating DataFrames

1) Add to the table a Boolean attribute called 'isContactKnown'. It should say for each element if the 'contact'

category is 'unknown' or is it 'cellular'/'telephone'

In [ ]:

2) Change the type of the 'campaign' attribute to Categorical data.

In [ ]:

Data Distribution

1) What is the range of values for each attribute in the table? what are the mean, std, median values for each category? Hint: You can do all this with one command!

In [ ]:

2) Plot a histogram of the 'balance' attribute and a boxplot of the 'age' attribute.

In [ ]:

3) Plot separately the distributions of balance values for people with negative outcome and positive outcomes. Are they different? If so, how?

In [ ]:

Q2

Missing Values

1) Which of the attributes have missing values? How many?

In [ ]:

2) Create another dataframe, which doesn't include any people with 'contact' which is 'cellular' or 'telephone'

In [ ]:

3) Look at the missing values in the 'days_from_last_contact' attribute. What do you think these missing values might represent? Can you find support in the table for your assumption? Hint: The attribute names have a meaning.

In [ ]:

Discretization

In class, you've seen three methods of data discretization, and we will focus on two of them - Equal-width (distance) and Equal-depth (frequency).

1) Discretize the balance attribute to 7 intervals using equal-width discretization, and plot the frequency table for them.

In [ ]:

2) Discretize the age attribute into 6 intervals using equal-depth discretization, and plot the frequency table for them.

In [ ]:

3) Propose a way to further discretize the Month column.

Type *Markdown* and LaTeX: α^2

Q3

Correlation

1) Calculate the pearson and spearman correlation between age and balance.

In [ ]:

2) Calculate the full pearson correlation matrix for all the numeric columns in the data.

In [ ]:

Mutual Information

Calcalate the entropy of each of the attributes in the data. Which has the biggest entropy?

In [ ]:

Which attribute has the highest Mutual Information with the 'outcome' attribute?

In [ ]:

Q4

Outliers

Does the campaign column in the table have any outliers? Support your claim with a relevant plot.

In [ ]:

Normalization

Think of an appropriate normalization method and normalize the days_from_last_contact column in the table. Why did you choose this method?

In [ ]: