CHURN ANALYSIS: FUNDRAISING

The data at hand is a part of the database of a fundraising organization. As always, before we can start the modelling phase, we need to **create the basetable**. In order to do this, the information of different tables has to be combined (see Appendix). The fundraising organization gave you a data dump of **March 2, 2020**. They will use the model you create today (**October 25**) to score their current customer base in terms of their **churn probability** (with today's data). A campaign will be launched and its results will be measured during the next full year.

1. Based on the above assignment, determine the **time windows** for both the purpose model as well as the model building phase. Store the relevant moments of your model building phase in the variables **end_independent**, **start_dependent** & **end_dependent**
2. Subset the *extrel* dataset according to the appropriate time window. Remember, we are trying to **predict** which donors will churn. This means that we are not interested in donors who already left or who are no donors yet. Thus, take into account the following:
   a. The **start** of the relationship should be **before (or equal to)** the end of the independent period.
   b. The **end** of the relationship should be later than the start of the dependent period (or missing)
3. Create the following independent variables. People with similar features are assumed to show similar behaviour. Different kinds of features are used. In classical churn models, there are two main types of features: behavioural features and demographics. Three predictors have shown to be key in predicting customer churn in previous work: recency, frequency and monetary value. When we think back at our purpose model, we can only use information that is available at the end of our independent period. Therefore, make sure you use only information available at the end of the **independent** period :
   a. Frequency: how often a donor has donated during the independent period.
   b. Recency: time (in days) since the last donation
   c. Total and average donation per donor (monetary value)
   d. Paytype per customer
      i. Create new variables that signify whether a donor ever used sendout, order, own initiative and unknown
   e. Preferred mailing language
   f. Dummy whether the donor ever uttered a complaint (CLASCODE)
   g. Dummy whether communication direction was ever incoming (CONTDIREC)
4. Create your **dependent variable** (do this for a partial churn and a complete churn)
5. **Merge** everything (using partial churn), make sure to include only those donors in the time window
6. You will see that at this point (depending on your merge), a great deal of observations are missing. **Impute** (replace) those **missing values** with a zero, but create an **additional variable** that has to have the value of 1 if a specific observation had a missing value and 0 otherwise.
7. Create a **full regression model** and **interpret** the results (note that we did not check any assumptions so interpretations might not be fully correct).

It is always advised that you draw an ERD before starting so that you know how the different tables are linked to each other. After that you can indicate which of these tables you need for your analyses and which you can discard.

8. Asses the **performance** of your model if we assume to have a budget to contact 20% of our current customers (20% most likely to churn). The needed measures are given in the notebook.

APPENDIX

Extrel: All the donors of the organization

Variable	Description
Extrelno	Unique identifier of each donor
Exrelactcd	Activity code of the donor
Extrelstdt	Start date of the relationship
Exreldaten	End date of the relationship (Missing: not ended)

Extrelty: Description of the activity

Variable	Description
Exrelactcd	Activity code of the donor
Exrelactde	Description of the activity

Nameaddr: Socio-demographical information

Variable	Description
Extrelno	Unique identifier of each donor
Name1title	Title to address someone
Postcode	Postcode
Languagecode	Preferred mailing language

Payhistory: Payment history of each donor

Variable	Description
Pid	Unique identifier for each payment
Pdate	Date of payment
Pamt	Amount of payment
Extrelno	Unique identifier of each donor
Paytypecd	Paytype *O Bank transfer* *D Permanent order* *E Own initiative* *X Unknown*
Status	Status of payment *OK Normal/Real payment* *CO Correction (internal)* *RF RF (Refund)* *RC Recall*

It is always advised that you draw an ERD before starting so that you know how the different tables are linked to each other. After that you can indicate which of these tables you need for your analyses and which you can discard.

Communication: All possible communication between the donor and the organization

Variable	Description
Contid	Unique identifier for each contact
Mediumcode	Medium of the contact (CI is unknown)
Mntopcode	Main topic code of the contact
Classcode	Class of the contact
Extrelno	Unique identifier for each donor
Contdirec	Direction of the communication *I Incoming* *P Outgoing*
Contdate	Date of the contact

Commediu: Description of medium type

Variable	Description
Mediumcode	Code of the mediumtype
Mediumdesc	Description

Commaint: Description of the main topic code

Variable	Description
Mntopcode	Main topic code
Mntopdesc	Description

Comclas: Description of the contact class

Variable	Description
Clascode	Code of contact class
Clasdesc	Description

It is always advised that you draw an ERD before starting so that you know how the different tables are linked to each other. After that you can indicate which of these tables you need for your analyses and which you can discard.