

Analysis of portfolio changes

You are newly hired as an analyst in an insurance company. The head of the analytics department needs more insight about the **changes in the portfolio the last 6 months**. You are provided with a dataset that gives information about each customer in the portfolio. Moreover, it gives also information about the customer at two certain timepoints, which we denote as TIME1 and TIME2. The time window between these two points in time is exactly 6 months.

There are particularly three type of changes/movements the department head is interested in:

1. Full churn. The manager asks for insight into the characteristics of those who choose to cancel their insurance contracts. **Full churners are identified in the data as having missing TIME2 features.**
2. Partial churn. Similarly, she asks you to provide insight about those customers who scale down the portfolio. **Partial churners are identified as those customers who have decreased their number of covers from TIME1 to TIME2.**
3. Moresale or cross sale. What characterize customers who choose to increase their portfolio of insurance products? **"Moresale customers" are identified as those customers who have increased their number of covers from TIME1 to TIME2.**

You are asked to build a notebook in either Python or R that:

1. Provides descriptive statistics of full churn, partial churn and moresale rates.
2. Provides full churn/partial churn/moresale rates broken down into interesting dimensions like tenure, age, size of portfolio etc. Motivate your choices about what is interesting and what is not interesting, seen from the perspective of the decision maker.

After the report is delivered your manager asks you to build prediction models for both full churn, partial churn and moresale.

1. Build one prediction model per target. You can choose whatever prediction technique you want. Important that you justify and explain your choice.
2. Provide a high-level explanation of the design of the models in the report and assess the overall predictive performance. Keywords here: ROC curve, Confusion matrixes, precision & recall & lift curves.
3. The dataset includes a feature called AVERAGE_INCOME_COUNTY_TIME1. Build models first without this feature. Make then a version **of the full churn model** where you include this feature. Describe the challenges you encounter when you use this variable and, moreover, show how you solve the problems that occur when including this feature.
4. Reflect upon what kind of features you miss in your data and explain why these are potentially important.
5. For both developers and consumers of ML-models it can be challenging to understand what the model has learned from the training data. Provide a brief discussion on interpretability to your model, and how this relates to the targets (full churn, partial churn and moresale). Shapley Values may be interesting to use here.

The models you have built are approved by your manager and she asks you to bring them into production.

Assignment BAN427: Insurance Analytics.
Submission deadline: September 3rd 2021 .

1. Describe how you will bring these models into production and particularly how the models optimally should support case handlers working in the retention/holdback team and case handlers in sales teams.
2. Furthermore, describe how you will test the effectiveness of these models in a production phase.

File description for dataset:

TIME1 =	“first date observed”
TIME2 =	“last date observed”
TENURE_TIME1=	“The tenure of the customer at time 1”
TENURE_TIME2 =	“The tenure of the customer at time 2”
AGE =	“The customers age at time 1”
WOMAN =	“Indicator for women. 1 if woman, 0 if man”
NUMBER_COVERS_TIME1 =	“The number of insurance covers at time 1”
NUMBER_COVERS_TIME2 =	“The number of insurance covers at time 2”
TOTAL_PREM_TIME1=	“The total insurance premium at time 1”
TOTAL_PREM_TIME2=	“The total insurance premium at time 1”
HAVE_CAR_AT_TIME1=	“Indicator for having car insurance at time 1, 1=have car”
CHILD_INSURANCE_AT_TIME1=	“Indicator for having child insurance at time 1, 1=have child”
TRAVEL_INSURANCE_AT_TIME1=	“Indicator for having car insurance at time 1, 1= have travel”
HOUSE_INSURANCE_AT_TIME1=	“Indicator for having car insurance at time 1, 1=have house”
CLAIM_EVENT_BEFORE_TIME1=	“Indicator for having experienced a claim event before time 1, 1 =yes”
PARTNER_CUSTOMER_TIME1=	“Indicator for being a partner customer, 1 =yes”
AVERAGE_INCOME_COUNTY_TIME1=	“An estimate of the average income in the municipality”