ERHVERVS AKADEMI AARHUS

Exercise 1.

In groups of 2-3.

What Is Churn Rate? *The churn rate, also known as the rate of attrition or customer churn, is the rate at which customers stop doing business with an entity.*

It is most commonly expressed as the percentage of service subscribers who discontinue their subscriptions within a given time period. It is also the rate at which employees leave their jobs within a certain period. For a company to expand its clientele, its growth rate (measured by the number of new customers) must exceed its churn rate.

Churn rate for a higher education in Denmark.

In this exercise we will look at churn for students attending a higher education in Denmark. The data is fictive, generated by a Python program. But the distributions are based on actual churn rates.

Churn is a problem both for the individual student (who stop an education before receving a diploma) and for the institution facing problems with churn. So, it would be helpful, if it was possible to predict churn, well in advance, based on the available data.

Given a correct prediction, it would then be possible to take preventive actions in the form of councelling, assignment to studygroups, allocation of resources for support with homework etc.

Your task as a datascientist assigned to the problem:

Is it possible, to predict whether a student will complete an education or not, based on the available data? This is the problem that you are here asked to investigate.

The dataset (studentchurn) can be downloaded in .csv format from Canvas.The dataset can be imported into a text program or a .csv reader (E.g. Openoffice is good at displaying the data from .csv files). A file, "studentchurn.py", for reading the dataset into a pandas dataframe is also available on Canvas.

The data contains an id for each student, line for previously completed education, grade average, distance from home to school, and membership of studygroup (if available).

A) Getting to know the data and the problem.

What kind of machine learning problem are we talking about in this exercise? Be as precise as possible.How many different features (not counting labels (y-values) as a feature here) are there in this dataset? How many of these features would the Y-data (the labels) consist of?

B) Cleaning the data

You have to decide on which of the features for the X training dataset you want to use and then later what to do with missing data for the features you keep.

You also need to consider that some the data is not numerical data, so that data would need to be converted into numbers (in both training and test set). This is easy to do with pandas – here is an example: `xtrain['Sex'] = xtrain['Sex'].replace(['female'],1.0 )`

This changes all the 'female' labels into 1.0 instead of test for the sex (of course the male would also need a value).

After cleaning (and scaling your data – see slides and previous exercises in this course on how to do this) then you should split your data into two sets. One for training and then one for evaluating performance. How many rows would you put into the training set and how many would you put into the test set? Give reasons for your decision.

C) Choosing a model and doing training. Evaluate performance on test set.

Choose a model for doing training and explain your reason for this choice. Train the model using the training data. Evaluate perfomance. ***Can you predict churn, yes/no? How well?***