**Data preprocessing – textual data**

Customer Churn is one of the most important and challenging problems for businesses such as Credit Card companies, cable service providers, SASS and telecommunication companies worldwide. Even though it is not the most fun to look at, customer churn metrics can help businesses improve customer retention.

there is the involuntary churn, for instance where a customer cannot pay their credit card bill and no longer stays with the credit card company.

* Drag the “customer churn data” into your process window (data can be found at training resources->data->customer churn data)
* Try to apply any classifier “ex. KNN” to classify churn issues, will it work?
* What are the proposed data pre-processing steps that you recommend applying to solve this classification problem?

**Data preprocessing – Image data (Discussion)**

We are looking to implement car license plate detection to the image below, what are the preprocessing steps that you recommend applying to make the task easier and more accurate?



**Feature Correlation and Feature dimensionality reduction**

Download the data set from the following link

<https://drive.google.com/file/d/1wOmkVs2zz4mbol-Y2gEGlv0N9y62Kz5J/view?usp=sharing>

Import the heating oil consumption.csv data set into your RapidMiner data repository and save it with a name.

* Drag the data set into your main process window. Run the data set to get a better understanding.
* Apply the Correlation Matrix operator.
* Which attributes have the strongest correlation? What do the correlation values ”Insulation rating - Heating oil use” and ”Insulation rating - Temperature” tell us?

Now you are going to use the Sonar dataset, where there are 50 frequencies of sonar readings as features, and the target variable is whether the reading was from a metal cylinder or a roughly cylindrical rock. Retrieve the Sonar data set from the Repository panel of RapidMiner and connect it to the output. Run the process. How many attributes does the data set contain?

* Apply the Linear Discriminant Analysis (LDA) and run the process.
* What is LDA useful for?
* For what purpose can the resulted probabilities be used?