

## **Project Update - Lennart and Jesse**

### **Are there any changes to your problem description or the approach that you are taking?**

No. The basic approach is still the same. We will implement at least two classifiers for each prediction type (classification (claim?: TRUE/FALSE) and claim-amount (continuous))

We realised that relatively high error rates are unavoidable, due to the high uncertainty inherent to the issue (everyone can have a car crash). We try to get information about how well good classifiers perform on this or similar problems.

### **Which classification/clustering/regression algorithms are you going to use?**

Neural Network and Support Vector for classification problem. Depending on the performance Decision Tree/Forest including visualisation.

For continuous prediction (given claim? == TRUE, how high claim-amount) we will apply a Linear Regression. We will experiment with features, and functions on feature values.

### **How will you evaluate/compare them?**

For an insurance company the overall, difference between predicted cost and actual cost of customers is important. Punctual wrong classifications or individual outliers are less of concern.

While we will analyse the confusion matrix for the classification problem, the overall/average difference between predicted cost and actual cost in the continuous prediction problem is of higher importance to us.

### **What is your progress till now? Describe shortly, no need for figures or code.**

We uploaded the data, converted several feature values into integer values (e.g. certain classes), imputed the data-set, and normalised it.

We implemented a few classifiers with the sklearn library, but haven't tested their performance yet.