Data

Download:

* [train.csv](http://seppe.net/aa/assignment1/data/train.csv)
* [test.csv](http://seppe.net/aa/assignment1/data/test.csv)

The data set consists of car insurance claims of a Belgian insurer for retail policy-holders (i.e. not companies but individuals). The data was obtained in the years 2017 (train set) 2018 (test set). The test set does not contain the target, so you will need to split up the train set accordingly into your own validation and test set. The test set supplied here is to rank and assess your model on the competition leaderboard. The features were obtained at the start of the handling process of each claim (registration/intake/entry). The reason for this being that we want the model to be utilized on new cases as they come in, rather than at the very end. The target describes claims which were identified in the past as fraudulent.

Variables:

* claim\_id: claim identifier (anonymized). Only used to uniquely identify each claim
* fraud: target (N or Y) -- was the claim fraudulent? **Only available in train set**
* claim\_amount: amount paid out by the insurance company. Used to evaluate your model. **Only available in train set, as this amount is not known at claim registration, i.e. early in the process.**
* claim\_date\_registration: date when the claim was registered (YYYYMMDD)
* claim\_date\_occurrence: date when the accident happened (YYYYMMDD)
* claim\_time\_occurred: time when the accident happened (HHMM) -- contains many missing values
* claim\_postal\_code: location where the accident happened
* claim\_cause: categorical variable describing what caused the accident
* claim\_liable: whether the policy holder is to some degree liable for the accident (N or Y)
* claim\_num\_injured: number of injured parties
* claim\_num\_third\_parties: number of third parties involved (e.g. in the accident)
* claim\_num\_vehicles: number of vehicles involved
* claim\_police: whether a policy report was written up
* claim\_alcohol: result of the alcohol test executed on the driver (Positive or Negative). A missing value indicates that no test was taken
* claim\_language: language the claim was registered in (Dutch or French)
* claim\_vehicle\_id: the vehicle that forms the object of the claim (anonymized), i.e. the object that incurred the damage to be covered by the insurer
* claim\_vehicle\_brand: brand of the vehicle
* claim\_vehicle\_type: type of the vehicle (only cars and larger cars/vans) are considered here
* claim\_vehicle\_date\_inuse: date the vehicle was registered (YYYYMM)
* claim\_vehicle\_cyl, claim\_vehicle\_load, claim\_vehicle\_fuel\_type, claim\_vehicle\_power: additional properties describing the vehicle: cylinder contents, load capacity, fuel type and power
* policy\_holder\_id: policy holder (anonymized)
* policy\_holder\_postal\_code: location of the policy holder
* policy\_holder\_form: legal form or gender of the policy holder. Only M or F in this case as we only consider retail customers
* policy\_holder\_year\_birth: year of birth of the policy holder
* policy\_holder\_country: country of the policy holder (Belgium or Non-belgium)
* policy\_holder\_expert\_id: identifier of the expert appointed by the policy holder to analyze the accident or damage, if any
* driver\_id, driver\_postal\_code, driver\_form, driver\_year\_birth, driver\_country, driver\_expert\_id, driver\_injured, driver\_vehicle\_id: similar. driver\_injured indicates whether the driver was injured. Note that the driver can be different from the policy holder. vehicle\_id is supposed to be the same as claim\_vehicle\_id, but in rarely this is not the case (most likely due to data quality issues)
* third\_party\_1\_id, third\_party\_1\_postal\_code, third\_party\_1\_injured, third\_party\_1\_vehicle\_type, third\_party\_1\_form, third\_party\_1\_year\_birth, third\_party\_1\_country, third\_party\_1\_vehicle\_id, third\_party\_1\_expert\_id: similar, repeated for maximally three third parties. Note that third\_party\_1\_form can include C here (legal entity, company)
* repair\_id, repair\_postal\_code, repair\_form, repair\_year\_birth, repair\_country, repair\_sla: information about the repair shop which performed repairs on the damaged object. repair\_sla indicates whether the repair shop has a service level agreement with the insurance company
* policy\_date\_start: date when the policy contract was started (YYYYMM)
* policy\_date\_next\_expiry: next expiry date of the contract (YYYYMM)
* policy\_date\_last\_renewed: date when the contract was last renewed (YYYYMM)
* policy\_num\_changes: number of changes made so far to the contract (i.e. changing the coverage, premium, etc.) so far (up until date of registration)
* policy\_num\_claims: number of previous claims registered for the contract (up until the date of registration)
* policy\_premium\_100: premium in hundreds of euros (rounded)
* policy\_coverage\_1000: coverage in thousands of euros (rounded)
* policy\_coverage\_type: included types of coverage, expressed as a binary string containing 9 bits

**Missing values are represented as an empty string. Missing values can occur virtually everywhere due to the point in time when the data is extracted. The process is heavily based on manual data entry, so expect data quality issues.**