Predicting total payments in Swedish Motor Insurance

A Data Management Plan created using DMPonline

Creator: Aleksandar Sibincic

Affiliation: Other

Funder: European Commission (Horizon 2020)

Template: Horizon 2020 DMP

ORCID iD: 0000-0002-4844-6306

Project abstract:

The experiment is based on Swedish Motor Insurance data set which describes third party automobile insurance claims for the year 1977 in Sweden. The goal here was to analyze the data and to perform simple linear regression for the 2 points of interest in this dataset: number of claims (the frequency) and sum of payments (the severity). In order to do that, we had to split the dataset into training and test set, apply predictions in the test set according to a training and to come up with measures for our predictions

Last modified: 15-04-2020

Predicting total payments in Swedish Motor Insurance - Detailed DMP

1. Data summary

State the purpose of the data collection/generation

In this experiment we try to predict the sum of payments for the number of claims in Swedish Motor Insurance. For this experiment we use linear regression, make our own predictions and at the end we find a measurement to see how good our predictions are.

Explain the relation to the objectives of the project

Since we want to make linear regression on Swedish Motor Insurance, we use only publicly available data which is a data set from 1977.

Specify the types and formats of data generated/collected

The input dataset is in csv format consisting of 2182 samples and 7 dimensions (columns). It does not have any missing values. Dimensions represented in dataset are: Kilometres, Zone, Bonus, Make, Insured, Claims and Payment.

Our generated datasets are also all in csv format and consist of 2 columns which are of the particular interest for this project which are:

- number of claims
- total sum of payments

Specify if existing data is being re-used (if any)

This input dataset is publicly available and as such reused in many other experiments.
Also we generate 2 datasets which will be used for training. One training set and one test set in ratio 80-20. This will be later used to train our regression line and to make predictions.

Specify the origin of the data

Input dataset is used in Ewards W. Frees' book "Regression Modeling with Actuarial and Financial Applications"(Cambridge University Press 2010), Chapter 20.5 "Case Study: Swedish Automobile Claims". It is compiled by the Swedish Committee on the Analysis of Risk Premium in Motor Insurance, summarized in Hallin and Ingenbleek (1983) and Andrews and Herzberg (1985). It is publicly available on kaggle

State the expected size of the data (if known)

The whole csv file with 7 columns and 2182 samples is 49.58 KB large.

As output two datasets are used for training and test of the linear regression with a 80-20 ratio and one consists of our predictions. They are in standard .csv format consisting of two columns (feature and label). Five plots are in .png format and for the measurement we choose RMSE which is obtained from console. Output data are generated using Python and they are in sum 285 kB large.

Outline the data utility: to whom will it be useful

Our output data can be useful for others who want to perform other
regression strategies and to compare them with linear regression used in this project.

2.1 Making data findable, including provisions for metadata [FAIR data]

Outline the discoverability of data (metadata provision)

Metadata can be read from the documentation/metadata.xml file within the repository.

Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?

Yes. The repository containing input data, generated data including datasets and plots, as well as source code to reproduce them, is located here

Outline naming conventions used

Naming conventions used in this project can be found here. Basically, for the files generated we use following names X_Y_Z.extension where:

- X represents the identifier for the experiment (SwedishMotorInsurance)
- Y represents description for the file (e.g. TrainingSet, PredictionSet...)
- Z is author's last name (Sibincic)

Outline the approach towards search keyword

For the keywords, we chose the important ones that can in best way describe out data. As mandatory we use our origin of dataset (Swedish Motor Insurance), our objective (Linear Regression) and some specific description for a specific data (e.g. Training set, Prediction, Distribution...)

Outline the approach for clear versioning

We wil use this standard for versioning. First publicated version will be 1.0.0

Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how

Metadata file was created using the Dublin Core standard

2.2 Making data openly accessible [FAIR data]

Specify which data will be made openly available? If some data is kept closed provide rationale for doing so

Whole data described before including input and generated datasets, plots, documentation, source code will be openly available.

Specify how the data will be made available

The whole repostiory is available on GitHub under this link:
https://github.com/aleksandarsibincic/SwedishMotorInsurance_LinearRegression

Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?

For downloading data it is needed Web Browser or also use can use wget. To access data simple tools are needed which almost any modern computer has. All what is needed is spreadsheet viewer (e.g Microsoft Excel, LibreOffice Calc..), image viewer and text editor. To reproduce code you will need Python with version > 3 and with some extra libraries like pandas and numdy.

Specify where the data and associated metadata, documentation and code are deposited

As already stated, the whole repository can be found on GitHub under this link:
https://github.com/aleksandarsibincic/SwedishMotorInsurance_LinearRegression

Specify how access will be provided in case there are any restrictions

We will use different repository hosting provider in case some problems appear in future with GitHub

2.3 Making data interoperable [FAIR data]

Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.

The datasets are stored as csv files and plots as png. Since it is also possible to change this in Python code by replacing one or two lines, there is actually no restrictions here

Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?

Datasets only use float numbers and commas. It is a standard vocabulary

2.4 Increase data re-use (through clarifying licenses) [FAIR data]

Specify how the data will be licenced to permit the widest reuse possible

The data as well as the source code are licensed under the CC0 license as stated in the LICENSE.md file in the repository.

Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed

There is no any embargo on using the data

Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why

The whole data in the repository is free to use without any restrictions

Describe data quality assurance processes

Since this is a university project only, there is no special data quality assurance process performed here

Specify the length of time for which the data will remain re-usable

The data will remain reusable indefinitely

3. Allocation of resources

Estimate the costs for making your data FAIR. Describe how you intend to cover these costs

Since GitHub is free, there is no actual cost for storage of data.. Besides working hours of the author there is no actual costs

Clearly identify responsibilities for data management in your project

Data management actually are really only analyzing datasets, transformation for the new training and test set, creation of prediction set and a measurement for it.

Describe costs and potential value of long term preservation

The only thing that could happen although unlikely is that GitHub will no longer be free. But still we assume there will be other free repository hosting provider so there are no actual costs

4. Data security

Address data recovery as well as secure storage and transfer of sensitive data

Complete software will be in GitHub and they will have their own recovery strategies. The same is also for datasets uploaded on Zenodo. Also, the author will hold the original data in case they are needed again.

5. Ethical aspects

To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former

All data are anonymized already in the input set and there is general public license to reuse it.
Also we used CC0 license so there should be no any legal issues

6. Other

Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)

There are no other procedures since this is only a part of university project