Exam Assignment

Machine Learning (BSc Data Science) Fall 2025
IT University of Copenhagen

1 Introduction and formalities

This is the project description for the exam project in the Machine Learning course for the BSc program in Data Science at the IT University of Copenhagen. The project must be submitted electronically via LearnIT no later than **December 18**.

Groups. Please register as a group in LearnIT by the end of the day on October 8. The project is designed for groups of 3 students. Therefore it is strongly recommended to work in groups of 2–3 persons. If you have any questions regarding the group formations, you need to talk to the course manager before the group formation deadline.

The course manager reserves the right to modify the grouping if necessary. If you need help finding a group, contact the course manager before the group formation deadline. Any students that have not registered as part of a group will be allocated to a group (possibly an existing one) at random.

Only one person for each group should submit the project.

1.1 Partnership with Tryg

The Machine Learning project investigates risk in insurance, and we are happy to announce that the project has been designed in partnership with the insurance company Tryg's Data Science group. The data are completely public, but the questions you will pursue are similar to those an insurance company data scientist might pursue.

The project is a normal university project in that it is evaluated by the course manager, according to criteria laid out in the course description.

The project has a special addition, though: The course manager will pick a couple of particularly good submissions, and share them with representatives from Tryg's data science team. From these nominated project, Tryg's team will pick a winning project. The authors of the winning project will receive a prize.

In case you do not want to consent to us sharing your project with the external partners at Tryg (which we will only do if we deem your project to be among the very best ones), you must opt out of the competition when you submit the project via LearnIT (opt-out happens in the submission portal).

1.2 What should be handed in?

You must hand in both a report and the source code you have developed during the project compressed as a single `.zip` file. Note that the project's evaluation will be based almost entirely on the report; the source code should be seen as a supporting document.

Make sure to use correct references in your report, including references to any of the course textbooks. This also applies to code; if you copy or take inspiration from code developed by others, this should be stated clearly in your report. This goes for **Generative Artificial Intelligence** (GenAI), too. You can use GenAI as an aid in developing code, but do not copy-paste, do make sure to give references, and remember that *you* are the expert. Please read and comply with the university's guide on using GenAI[1]. Note that any work based on a previous exam submission for this course, even if it is your own work, needs to be clearly marked as such and cited.

Report. The report should be submitted as a single PDF file. There is a strict limit of 10 pages, including figures, tables, code snippets, references, and appendixes, but excluding the front page. The

[1] `https://itustudent.itu.dk/Study-Administration/Generative-AI`

project must be typeset with at least 11pt font size and margins of at least 2 cms. The report must be in PDF format and have a front page that meets the ITU requirements.[2]

Implementation and code. Your implementation has to be in Python. Except where explicitly stated, there are no restrictions as to which Python libraries you may use.

Your code should be organized such that it is easy to read, i.e. you have to use descriptive names for files, functions, variables, etc. The code may be organized in regular Python source files (`.py` files) or Jupyter notebooks (`.ipynb` files).

2 Problem and data set

In this project, you will explore different methods to model claims risk in insurance of automobiles. The dataset for the project contains an overview of vehicular insurance policies and their claims history over the last year. It consists of ∼68000 entries, where each entry describes all related information about a policy, represented within 12 columns as:

Name	Description
IDPol	The ID of the Policy, which is a unique identifier representing the contract of a person and their car to the insurer.
ClaimNb	The number of claims (i.e. count of third-party liability claims) that occurred during the exposure period for that policy.
Exposure	The amount of "time at risk" (in years) that the policy was active/observed. For instance, 1 means full year, 0.5 means half a year, etc.
VehBrand	The (anonymized) vehicle brand (or brand category). It is often coded (e.g. "B1", "B2", etc.) rather than the full car brand name.
VehGas	The fuel type of the car (e.g. "Regular" or "Diesel").
VehPower	The power rating of the vehicle. Higher values correspond to higher vehicle power.
VehAge	Age of the vehicle in years (how old the vehicle is).
DrivAge	Age of the driver (or principal driver) in years.
Area	The area (or density class) of the city / community where the car driver lives. This is an ordinal label starting from "A", where "A" is most rural, "B" is less rural, and so on.
Density	The number of inhabitants per square-kilometer in the location where the driver lives (i.e. population density).
Region	The administrative region in which the policy / driver resides.
BonusMalus	The "bonus/malus" rating (a French insurance concept), typically in a range (e.g. from 50 up to ∼350). If the value <100, that indicates a "bonus" (good record), while >100 indicates a "malus" (worse record).

The dataset is divided into a training set containing 80% of the entries (∼54400 claims) and a test set containing 20% (∼13600 claims), available in .csv format as:

`claims_train.csv` and `claims_test.csv`

[2]Found at: `https://itustudent.itu.dk/study-administration/exams/submitting-written-work`

3 Scientific requirements to the project

As this project aims to investigate methods for determining claims risk of insurance customers, it is up to you to define a reasonable measure for claims risk given the data (see below). You should carry out and report on the investigations of claims risk making sure that you cover all of the tasks set out below.

3.1 Data Cleaning, Exploratory Data Analysis and Visualization

You should carry out a cleanup and analysis of the claims dataset, observing all features to note whether any of them go outside of expected ranges (e.g. Exposure), then illustrate selected aspects of the data. You should also present a visualization of the dataset based on both dimensionality reduction using Principal Component Analysis (PCA) and a Clustering method.

You are recommended to consider applying some kind of feature scaling to at least one feature as part of your analyses of the data.

3.2 Methods

You should explore at least three machine learning methods:

M1. Decision Tree regressor with at least one categorical variable.

M2. Feed-Forward Neural Network regressor.

M3. One or more other methods of your own choice, not limited to regression methods.

The first two methods, M1, and M2 should be implemented in two versions:

1. An implementation from scratch (training and prediction) using only Python standard libraries and the numerical libraries NumPy and SciPy.

   Thus, you cannot use machine learning libraries such as TensorFlow, PyTorch, or Keras. Note that the restriction only applies to the implementation of the method itself (training and prediction), but not any pre-processing before you feed the input to the method or the further interpretations of the results, such as visualizations.

2. A "reference implementation" using any Python library, which can assist in asserting the correctness of your own implementation.

For the third method (M3) you may use any library you wish with no restrictions.

3.3 Report

The report should contain the following things.

Definition of target variable(s) and evaluation metrics. You should argue for the definition of "claims risk" you choose to treat as your target variable. In your report, you should also define and argue for the metrics you will use to evaluate the performance of your models.

Data Cleaning and Exploratory Data Analysis. Your report should introduce the reader to the data by illustrating selected aspects of the data, and argue which features needed cleaning and reason for doing so.

Description of feature extraction and selection. If you choose to coarse grain certain features, combine features, disregard features, etc., your report should explain how you do this and why.

Visualization of data. As part of the exploratory data analysis, your report should present a visualization of the dataset based on Principle Component Analysis and a clustering method (not necessarily in combination). Remember to state what the reader can learn from the visualization.

Details on implementations. Your report should describe and discuss the key points of how you implemented methods M1, M2, and M3. For each method please make sure to include:

- A description of how you applied the method to the data, including details needed for an independent reproduction of your results.
- A discussion on how you have gone about selecting any hyperparameters for the method.
- A discussion on how you have asserted your implementation's correctness (i.e. did it do as intended?). Do note that Correctness and Performance are not interchangeable terms.

Interpretation and discussion of the results. Your report should include a thorough discussion of the comparative performance of methods applied. In particular, you should compare each methods' performance and guide the reader in interpreting the results. Use your expert knowledge to explain the results; for instance, why do particular methods perform better or worse than others?

4 Final comment

Modeling and predicting claims risk in insurance is a difficult problem. Hence, you should expect your machine learning models to do quite poorly. Your job is to do your best to improve the model's performance, and to make informed and reasonable decisions on every step of your modeling journey.

Have fun working on the project and I look forward to seeing your reports.