**CAS ETH ML in Finance and Insurance**

**BLOCK I: Intro to Machine Learning**

**Project “Claims frequency model”: Group Report**

Authors: Andreas, Tomasz, Francesco

**Note on the report**

1. The report is max 3000 words long (<10 pages), considering its main content and excluding References (if any). Its minimum length is ~1500 words (<5 pages), considering its main content and excluding References (if any). Please use a 12pt font (e.g., Arial or Times New Romans), with single or 1.5 spacing.

2. **Please delete all content in yellow and the questions before submission**. The report has to be shared in .pdf format.

**1. Business Context**

1.1 Project motivation

The aim of this project is to explore the feasibility and assess added value of using deep learning methodology to improve claim frequency modelling on a motor portfolio.

1.2 Knowledge gap/opportunity

Given the increasing amount of data available and relatively cheap computing power we want to pilot a neural network to estimate claims frequency and compare the results with traditional GLM modelling.

The projects should help us to build the knowledge required to develop internally deep learning model. Actuaries are using GLM models as they are part of a consolidated process and because of its relative easiness of interpretability that allows a seamless integration with severity models used in pricing.

Deep learning however should help us to understand with a higher level of precision the interaction between policy features and claim frequency with less distortion due to correlations or limitations in the number of variables to be considered in the regression model.

Better understanding of claims frequency drivers could be used to improve pricing models directly by upgrading the existing models.

Constraints with the integration with severity models and implementation in IT systems calculating the policy premium need to be addressed. We believe however that the knowledge resulting from the deep learning modelling can be used to improve the traditional claims frequency models. This can be achieved through feature engineering and better selection of variables to be used in the traditional GLM model.

**2. Problem Formulation**

2.1 Modeling challenge

Build a claims frequency model using deep learning and compare it with a traditional GLM model.

Both models will be built and optimized to compare results, weaknesses and advantages.

2.2 describe available data

A large market dataset from the French motor market is publicly available and is used for this project. The dataset is representative of a large portfolio but offers a relatively small number of features. We believe its size and complexity is adequate for the purpose of this project aimed to compare different modelling methodologies.

**3. Machine Learning Pipeline**

Describe the ML pipeline implemented in the Python notebook by summarizing the following items:

3.1 import of resources

3.2 data generation

3.3 data exploration, e.g., histograms, scatterplots, correlations

3.4 train vs. test data or other sampling methods

3.5 modeling:

* model classes
* hyperparameters, including regularization
* performance measures
* performance on test data: results
* interpretability of modeling results

**4. Business Application of the ML Models**

4.1 Model application

Claims frequency modelling has multiple possible applications in the insurance sector. In principle all areas where the number of events is object of the prediction, the model could be leveraged.

Our main focus in the project is around estimating claims frequency for pricing purposes but this definitely not the only use case.

* Pricing: claims frequency is building block of pricing where it is usually combined with a severity model to estimate the expected cost per policyholder. Best estimate of the expected cost is then adjusted (prudency margin, cost of capital, target profitability,…) to get to the tariff premium.
* Modelling numbers. This is the general scope of the model and fits into many uses cases. From estimating the number of calls the claims call canter will receive to the number of policyholders leaving the Company following a premium increase (elasticity of demand).
* Business intelligence. Regardless of what is the specific use case, the results of the model can be used to better understand the underlying process and learn about it. The deeper understanding of the dependency model and the contribution of each feature to the total expected number can be used to support business strategy and decision making. The predictive power is not the only use case of the model.
* Claims reserving. Where claims reserving methods follow a frequency severity approach or a projection of a claims count triangle is needed, this model can be applied as an alternative to traditional approaches.

**5. Limitations and Future Work**

5.1 Modeling limitations

how to improve your ML modeling approach? (E.g., collecting more data, engineering additional features, improving the ML pipeline...)

The use of the French market database is valuable for prototyping and compare results across different methods. For a real world application more work is required on data. Data from the Company own portfolio are needed to ensure applicability of the results:

* more variables should be available on the claims characteristics and on the policyholder.
* more realistic variables (e.g.: bonus malus is not a standard in all markets) should be available depending on the product and market

Feature engineering is another area of limitation of the current project:

* additional policyholder information could be very useful as it could allow to create new features on time series or across policies with the same policyholder. Behavioural features or features representing client “value” could be predictive and innovative in pricing models.

5.2 Business application limitations

how to improve the business application of your ML modeling approach?

Traditional pricing frequency severity models are a consolidated practice in the industry. The key limitation of new techniques is their interpretability and integration with existing methodologies and processes.

Neural networks are no exception as while their predicting power is one step forward compared to classic regression models, there is no clear meaning of the estimated parameters and their relationship with the features. This is a major limitation for explainability of the tariff to the stakeholders. In principle it should be possible to present a tariff in a simple and intuitive way (i.e.: area B gets a higher coefficient than area A as it is riskier based on statistics).

A second limitation is given by IT systems needed to calculate premiums that are in general able to cope with linear combinations of inputs (age, area,…) and weights (tariff). The calculation of a premium based on a neural network may require an investment that should be paid off in a short time frame (i.e.: extra margins due to better pricing must be proved).

Traditional products pricing however is not the only use case for number/frequency modelling therefore once the models are available and it is clear how to train them it should be possible to create value on the use cases where traditional regression models are not traditionally used. Business intelligence to support decision making or help procing and reserving teams with insight to be used in traditional models (e.g.: feature engineering for pricing or early detection of trends in reserving).