

A comparative Analysis Between Various Machine Learning Models and Generalized Linear Models

Jan Mikael Yousif*

February 2023

Abstract

In recent decades there has been a vast improvement in computational power which has lead to an increased demand for advanced modelling techniques such as Gradient Boosting Machines and Neural Networks. This thesis studies if the mentioned models have the ability to predict the claim frequency for an insurance portfolio more accurately than a traditional Generalized Linear Model (GLM). By training and Cross-Validating the mentioned models the thesis shows that the Machine- Learning models do perform better than the GLM for a data-source from a real insurance portfolio. The improvements that the Machine-Learning models resulted in were initially expected to be of greater magnitude but showed to only have a slight difference from the GLM. This is largely explained by the GLM already being a good predictor for the specified data. It is also observed that further advantages can be obtained by partitioning the data by the feature levels where the models performs the best. By training each model on the partitioned data, the weaknesses in each model are minimized while the strengths are more highlighted. Mixing models across partitioned data however leads to a large cost in interpretability which is often an important factor to consider when building price strategies in real-world applications.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden.
E-mail: janyousif@gmail.com. Supervisor: Filip Lindskog.