CATEGORICAL PREDICTOR EMBEDDINGS IN ACTUARIAL MODELING

A PREPRINT

Kevin Kuo
RStudio and Kasa AI

kevin@kasa.ai

Ronald Richman
QED Actuaries and Consultants

ronaldrichman@gmail.com

November 17, 2020

ABSTRACT

Enter the text of your abstract here.

***K*eywords** blah · blee · bloo · these are optional and can be removed

1 Introduction

Categorical data are modelled by actuaries in many different contexts, from pricing of insurance products to reserving for the liabilities generated by these products. Sometimes, these data are modelled in an explicit manner, for example, when building models that apply across multiple categories, a form of dummy coding is usually used. For example, when building models to price the frequency of motor insurance claims, claim experience relating to different types of motor vehicle will often be modelled by including a (single) factor within models that modifies the relative frequency predicted for each type of vehicle. In other cases, modelling is performed for each category separately, thus the categorical data is used within the modelling in an implicit manner, for example, common practice is to estimate reserves for different lines of business separately, meaning to say, with no parameters being shared across each of the reserving models.

Recently, several studies of insurance problems have applied an alternative approach, originally from the natural language processing literature (Bengio et al. 2003) and now applied to diverse types of machine learning problem (Guo and Berkhahn 2016), known as **categorical embeddings**. Instead of trying to capture the differences between categories using a single factor, the categorical embedding approach rather maps each category to a low-dimensional numeric vector, which is then used within the model as a new predictor variable.

This approach to modelling categorical data has several advantages over more traditional treatments of categorical data. Using categorical embeddings instead of traditional techniques has been shown to increase predictive accuracy of models, for example, see (Richman 2018) in the context of pricing. Models incorporating categorical embeddings can be pre-calibrated to traditional actuarial models, increasing the speed with which these models can be calibrated and leading to models with better explainability (Wüthrich and Merz 2019). Finally, the similarity between the vectors learned for different categories can be inspected, sometimes leading to insights into the workings of models, see, for example, (Kuo 2019).

On the other hand, several open questions about the use of embeddings within actuarial work remain, which we aim to address in this study. First, hyperparameter settings for embeddings, such as the dimensions of the embedding layer and the use of regularization techniques such as dropout or normalization, that achieve optimal predictive performance has not yet been studied in detail in the actuarial literature. In this work, we aim to study how embeddings using different settings perform in the context of a large-scale predictive modelling problem, and give guidance on the process that can be followed to determine this in other problems. Although neural network have been shown to achieve excellent predictive accuracy on actuarial tasks, many actuaries still prefer to use GLM models for pricing tasks, thus, the issue of whether transferring embeddings to GLM models can achieve better performance is considered in this paper. Traditional

actuarial techniques such as credibility theory have been used to work with some types of categorical data, but no study has been performed whether this can be applied within embeddings; here, we investigate whether performance is enhanced by applying credibility theory on embeddings relating to categorical variables with many labels. Whereas embedding layers are usually considered in the context of categorical data,the option exists to quantize numerical data and model it using embeddings, however, the results of doing so have not been investigated. Finally, in the past several years, a new type of neural network architecture based on attention (Vaswani et al. 2017) has been successfully used on embeddings in the field of natural language processing and we incorporate attention based models into our predictive modelling example.

In this work, we utilize the recently released National Flood Insurance Program (NFIP) dataset (Federal Emergency Management Agency 2019) which provides exposure information for policies written under the NFIP since XXXXDateXXXX, as well as the claims data relating to these exposures. We refer the reader to Appendix **??** for an exploratory analysis of this dataset.

The rest of this manuscript is organized as follows. Section 2 reviews recent applications of embeddings in the actuarial literature. Section 3 provides the notation used in the paper and defines GLMs, neural networks and related modelling concepts, including embeddings and attention. In Section **??**, we provide initial models for the NFIP dataset and consider the influence of hyperparameter choices on the results of the neural network model. Section 6 considers how successfully the embeddings used in the neural network model can be transferred to a GLM model of the same data. Extensions to the modelling approach are considered in Section 7 and we focus on attention based models in Section 3.2.2. The interpretability of embedding layers is addressed in Section 5. Finally, Section 9 provides a discussion of the results of this paper and considers avenues for future research.

(TODO: link to github repo)

2 Literature Review

Categorical data are usually modelled within GLMs and other predictive models using indicator variables which capture the effect of each level of the category, see, for example, Section 2 in (Goldburd et al. 2020), using one of two main encoding schemes: dummy-coding and one-hot encoding. Dummy-coding, used in the popular `R` statistical software, assigns one level of the category as a baseline, for which an indicator variable is not calibrated, and the rest of the levels are assigned indicator variables, thus, producing estimates within the model of how the effects of each level differ from the baselines. One-hot encoding, often used in machine learning, is similar to dummy-coding, but assigns indicator variables to each level, in other words, calibrates an extra indicator variable compared with dummy-coding.

A different approach to modelling categorical data is credibility theory (see (Bühlmann and Gisler 2005) for an overview), which, in the context of rating, can be applied to derive premiums that reflect the experience of a particular policyholder, by estimating premiums as a weighted average between the premium produced using the collective experience (i.e. of all policyholders) and the premium produced using the experience of the particular policyholder. The weight used in this average is called a credibility factor and is calculated with reference to te variability of the policyholder experience relative to the variability of the group experience. In this context, the implicit categorical variable is the policyholder under consideration.

Generalized Linear Mixed Models (GLMMs) are an extension of GLMs that are designed for modelling categorical data using a principle very similar to that of credibility theory (Klinker 2010). Instead of calibrating indicator variables for each level of the category, GLMMs estimate effects for each of these levels as a combination of the overall group mean and the experience in each level of the category.

Embedding layers represent a different approach to the problem of modelling categorical data that was recently introduced in an actuarial context. Note that in the next section, we reflect on similarities between the conventional approaches discussed above and embedding layers. (Richman 2018) reviewed the concept of embedding layers and connected the sharing of information across categories to the familiar concept of credibility theory. In that work, two applications of embedding layers were demonstrated. The first of these was in a Property and Casualty (P&C) pricing context, it was shown that the out-of-sample accuracy of a neural network trained to predict claims frequencies on motor third party liability was enhanced by modelling the categorical variables within this dataset using embedding layers. Second, a neural network with embedding layers was used to model all of the mortality rates in the Human Mortality Database, where the differences in population mortality across countries and the differences in mortality at different ages were modelled with embedding layers, again producing more accurate out of sample performance than the other models tested.

Contemporaneous with that work is the DeepTriangle model of (Kuo 2019), which applied recurrent neural networks to the problem of Incurred but not Reported (IBNR) loss reserving, to model jointly the paid and incurred losses in

the Schedule P dataset. Embedding layers were used to capture the effect of differences in reserving delays and loss ratios for each company in the Schedule P dataset. Evaluating the results of the DeepTriangle method showed that the out of sample performance of the model (tested against the lower triangles in the Schedule P dataset) exceeded that of traditional IBNR reserving techniques.

Many other applications of embeddings have subsequently appeared in the actuarial literature. Within mortality forecasting, (Richman and Wüthrich 2019) and (Perla et al. 2020) both apply embeddings layers to model and forecast mortality rates on a large scale. (Wüthrich and Merz 2019) discussed how embeddings can be calibrated using GLM techniques and then incorporated into a combined actuarial neural network, with subsequent contributions in P&C pricing by (Schelldorfer and Wüthrich 2019) and in IBNR reserving by (Gabrielli 2019) and (Gabrielli, Richman, and Wüthrich 2019). Other applications in IBNR reserving are in (Kuo 2020) and (Delong, Lindholm, and Wuthrich 2020) who use embedding layers to model individual claims development.

3 Definitions

In this study, we are concerned with regression modelling, which is the task of predicting an unknown outcome y on the basis of information about that outcome contained in predictor variables, or features, stored in a matrix X. For simplicity, we only consider the case of univariate outcomes, i.e., $y \in \mathbb{R}^1$. The outcomes and the rows of the predictor variable matrix are indexed by $i \in \{1 \dots I\}$, where i represents a particular observation of $(y_i, \mathbf{x}_i)$, where bold indicates that we are now dealing with a vector. The columns of the predictor variables are indexed by $j \in \{1 \dots J\}$, where j represents a particular predictor variable, of which J have been observed, thus, we use the notation X_j to represent the jth predictor variable and $X \in \mathbb{R}^J$. Formally, we look to build regression models that map from the predictor variables $\mathbf{x}_i$ to the outcome y using a function f of the form:

$$\mathbf{f} : \mathbb{R}^J \mapsto \mathbb{R}^1, \qquad \mathbf{x_i} \mapsto \mathbf{f}(\mathbf{x_i}) = y.$$

In this study, will use mainly use GLMs and neural networks to approximate the function $f(.)$.

The predictor variables that we consider here are comprised of two types: continuous variables, taking on numerical values and represented by the matrix X_{num} with J^{num} columns, and categorical variables, which take on discrete values indicating one of several possible categories, represented by the matrix X_{cat} with J^{cat} columns, such that $J^{num} + J^{cat} = J$.

3.1 Categorical data modeling

A categorical variable $X_j, j \in J_{cat}$ takes as its value only one of a finite number of labels. Let the set of labels be $\mathcal{P}^j = \{p_1^j, p_2^j, \dots, p_{n_{\mathcal{P}^j}}^j\}$, where $n_{\mathcal{P}}^j = |\mathcal{P}^j|$ is the cardinality or number of levels, in $\mathcal{P}^j$. One-hot encoding maps each value $x_{i,j}$ of X_j to $n_{\mathcal{P}}^j$ indicator variables, which take a value of 1 if the label of $x_{i,j}$ corresponds to the level of the indicator variable, and 0 otherwise. An example of one-hot encoding is shown in Table **??**.

Table 1: Example one-hot encoding of the state variable

state	state_CA	state_MD	state_ND	state_UT	state_WA
CA	1	0	0	0	0
MD	0	1	0	0	0
ND	0	0	1	0	0
UT	0	0	0	1	0
WA	0	0	0	0	1

One-hot encoding is often used in the machine learning community while the statistical community often favors dummy coding, which, instead of assigning n_p^j indicator variables, assigns one of the levels of the categories as a baseline, and maps all of the other $n_p^j - 1$ variables to indicator variables. An example of dummy encoding is shown in Table **??**.

Table 2: Example dummy encoding of the state variable

state	state_MD	state_ND	state_UT	state_WA
CA	0	0	0	0
MD	1	0	0	0
ND	0	1	0	0
UT	0	0	1	0
WA	0	0	0	1

After encoding the categorical data in this manner, most regression models such as GLMs will then fit coefficients for each level of the category in the table (if a tree based model is used, such as decision tree, then splits in the tree may occur depending on the presence, or not, of the categorical variable for the data). If one-hot encoding has been used, $n^j_{\mathcal{P}}$ coefficients will be fit, compared to $n^j_{\mathcal{P}} - 1$ coefficients in the case of dummy coding.

These coefficients represent the effect that each level of the categorical variable will have on the outcome. In the case that there are no other variables available in the dataset, then the coefficients will reflect the average value of the outcomes for that level of the categorical variable. For example, suppose that the categorical variable is a policyholder identifier, and the outcomes are the value of claims in different years, then the coefficients will reflect the average annual claims for each policyholder based on the experience. In other words, both of these encoding schemes give full credibility to the data available for each category, thus, even if a relatively small amount of data is available for a specific policyholder, the coefficient that is calibrated will only reflect that data. On the other hand, a foundational technique within actuarial work is the application of credibility methods, which are used for experience rating and other applications. These techniques provide an estimate that reflects not only the experience of the individual policyholder but also that of the collective, based on an estimate of how credible the data for each individual is. While we have described the application of credibility in a simple univariate context, it is also possible to apply credibility considerations within GLMs. using models known as Generalized Linear Mixed Models or GLMMs, and we refer to (Klinker 2010) for more details.

Having described traditional approaches for modeling categorical data, we now turn to neural networks, and discuss embedding layers for categorical data modeling, which we define in more detail in the section on neural networks.

3.2 Neural Networks

Neural networks are flexible machine learning models that have recently been applied to a number of problems with Property and Casualty (P&C) insurance. Here, we provide a brief overview of these models, and refer the reader to (Richman 2018) for a more detailed overview. Neural networks are characterized by multiple layers of non-linear regression functions that are used to learn a new representation of the data input to the network that is then used to make predictions. Here we focus on the most common type of neural networks, which are fully connected networks (FCNs), which provide as the output of each set of non-linear functions to the subsequent layer of functions. Formally, a K-layer neural network is:

$$\begin{aligned} z^1 &= \sigma(a_1.X + b_1) \\ z^2 &= \sigma(a_2.z^1 + b_2) \\ &\vdots \\ z^K &= \sigma(a_K.z^{K-1} + b_K) \\ \hat{y} &= \sigma(a_{K+1}.z^K + b_{K+1}), \end{aligned} \tag{1}$$

where the regression parameters (weights) for each layer

$$k \in [1; K]$$

are represented by the matrices a_k and the intercept terms are represented by b_k. Whereas the calculation inside each of the layers is nothing more than linear regression, σ represents the non-linear part of each layer. Choices for σ are often the tanh function or the rectified linear unit (ReLu) $max(0, x)$. The parameters of the network are estimated ('trained') as follows. First a loss function $L(.,.)$ is specified for the network that measures the difference between the observed data y and the predictions of the network $\hat{y}$, for example, the Mean Squared Error ($(y - \hat{y})^2$). Then, the parameters of the network are changed such that the loss decreases (formally, this is done using the technique of

backpropagation).Finally, training is stopped once the predictive performance of the network on unseen data is suitably good.

If K is set equal to 1, then Equation 1 reduced to nothing more than a GLM. A neural network with $K = 2$ is called a shallow neural network and for $K \geq 2$, the network is called a deep neural network. The matrix X of data input to the network can be composed of both continuous variables as well as categorical variables, which can be pre-processed using one-hot or dummy encoding. As mentioned above, a different option is to use encodings, which we discuss in more detail next.

3.2.1 Embeddings

Common issues with the traditional encoding schemes for categorical data occur when the number of levels for each variable is very large. Often, in these cases, models do not converge quickly, and the very large matrices that result from applying these schemes often cause computational difficulties. Besides for these practical issues, a deeper issue is that one-hot or dummy encoded data assumes that each category is entirely independent of the rest of the categories, in other words, there are no similarities between categories that could enable more robust estimation of models. In technical terms, this is because the columns of the matrices created by one-hot encoding are all orthogonal to each other. (These arguments appear in a similar form in (Guo and Berkhahn 2016).) Solutions to these problems are provided by embedding layers.

An embedding layer is a neural network component which maps each level of the categorical data to a low dimensional vector of parameters that is learned together with the rest of the GLM or neural network that is used for the modeling problem. Formally, an embedding is

$$z_{\mathcal{P}^|} : \mathcal{P}^| \to \mathbb{R}^{q_{\mathcal{P}^|}}, \qquad p^j \mapsto z_{\mathcal{P}^|}(p),$$

where $q_{\mathcal{P}^|}$ is the dimension of the embedding for the jth categorical variable and $z_{\mathcal{P}^|}(.)$ is an function that maps from the particular element of the labels p to the embedding space. Equation **??** states that an embedding maps a level of a categrical variable to a numerical vector. This function is left implicit, meaning to say, we allow the embeddings to be derived during the process of fitting the model and do not attempt to specify exactly how the embeddings can be derived from the input data. In Table **??** we show an example of two dimensional embeddings for the state variables, where these have been generated randomly.

Table 3: Example (random) embeddings of the state variable

state	dimension1	dimension2
CO	1.5115220	2.2866454
DC	-0.0946590	-1.3888607
ME	2.0184237	-0.2787888
PA	-0.0627141	-0.1333213
UT	1.3048697	0.6359504

When applying embeddings in a data modeling context using neural networks, the values of the embeddings will be calibrated during the same fitting process that calibrates the parameters of the neural network.

3.2.2 Attention

4 Predictive Modeling with GLM and a Minimalist Neural Network

In this section, we describe, fit, and evaluate both a minimalist neural network architecture utilizing embeddings and a traditional GLM. The working example for our experiments is as follows: Given a set of claims characteristics, we predict the losses paid on the property coverage of the policy. In the rest of this section, we describe the dataset we use, describe formally the models being considered, and discuss results.

4.1 NFIP data

The data we use comes from the National Flood Insurance Program (NFIP) and is made available by the OpenFEMA initiative of the Federal Emergency Management Agency (FEMA) (TODO: cite). Two datasets are made available by

OpenFEMA: A policies dataset with exposure information, and a claims dataset with claims transactions, including paid amounts. Because there is no way to associate records of the two datasets, we are limited to fitting severity models on the claims dataset. While the complete dataset contains over two million transactions, for the purposes of our experiments we limit ourselves to data from 2000 to August 2019, which amounts to approximately 1.4 million claims. The dataset can be downloaded from Cellar (TODO: link). The dataset includes a rich variety of variables, from occupancy type to coarse coordinates. For our models, we work with a few selected variables that represent continuous and discrete variables of low and high cardinalities, which we list in Table 4.

Table 4: Variables used in modeling.

Variable	Type
Building insurance coverage	Numeric
Basement enclosure type	Categorical
Number of floors in the insured building	Categorical (binned in original dataset)
Flood zone	Categorical (high cardinality)
Primary residence	Categorical (binary indicator)

4.2 Severity GLM

We fit a GLM with the following specifications:

- Target variable: Amount paid on building claim
- Predictors: *Log* of **building insurance coverage**, **basement enclosure type**, **number of floors in the insured building**, *prefix* of **flood zone**, and ***primary residence***.
- Link function: log
- Distribution: gamma

We take the log of the continuous predictor **building insurance coverage** following (cite cas glm monograph), which allows the scale of the predictor to match that of the target variable. Because the **flood zone** variable in the original data contains 60 levels, we take the prefix of the zone code, which corresponds to the level of risk as determined by FEMA (source?). For example, *A01*, *A02*, and so on are recoded as simply *A*.

A log link together with the gamma distribution is a standard choice for severity modeling, which provides a multiplicative structure where the response is positive.

4.3 A Simple Neural Network

(Note: this section is WIP doesn't match up w/ code, which is also WIP) For the neural net, we utilize the same responsible variable, amount paid on building claim. We normalize the numeric variable ***building insurance coverage***, one-hot encode the **primary residence** variable, and apply embedding layers of dimension one to the rest of the categorical variables. We then connect the embeddings and the normalized variable to a single output unit with softplus activation, where the softplus function is defined to be

\begin{equation} x \mapsto \log(1 + e^x). \end{equation}

Figure (todo: add figure) exhibits the architecture of our simple neural network.

4.4 Performance evaluation framework and results

To evaluate these baseline models and other models in this paper, we perform 10-fold cross validation and compare the root mean square error (RMSE) on the predicted and actual paid amounts.

5 Interpretations

Extracting and interpreting trained embeddings, including visualization techniques Can we explain the learned embeddings using a GLM? distance to water etc Rainfall Density Distance to coast

6 Transfer Learning

How can embedding layers be used in GLM models? (Potential uses of embedding layers as a feature engineering technique for GLM) Transfer learned embeddings to GLM - backprop on NN and use as preprocessing Train GLM on embedding layers - backprop on 1 layer NN

7 Extending the model

Investigate approaches for using embedding layers for numeric variables, or numeric variables that were captured as categorical variables (Can the modelling of numerical variables benefit from the application of embedding layers?) How should this best be done? CatBoost - cut into 256 groups smoothness Map numeric to a dense layer and use that as embedding - basis expansion Add a distribution to the embeddings Can insights from credibility theory lead to enhanced embeddings? Take credibility mixture over embeddings Cluster embeddings PCA of embeddings

8 Attention based modelling

Can we get some lift by adding attention layers? TabNet model Transformer model

9 Conclusions

Conclusions

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. "A Neural Probabilistic Language Model." *Journal of Machine Learning Research* 3 (6): 1137–55. https://doi.org/10.1162/153244303322533223.

Bühlmann, Hans, and Alois Gisler. 2005. *A Course in Credibility Theory and its Applications*. Springer Science & Business Media. https://doi.org/10.1080/03461230600889660.

Delong, Lukasz, Mathias Lindholm, and Mario V. Wuthrich. 2020. "Collective Reserving using Individual Claims Data." *SSRN Electronic Journal*, May. https://doi.org/10.2139/ssrn.3582398.

Federal Emergency Management Agency. 2019. "FIMA NFIP Redacted Claims Data Set." https://www.fema.gov/media-library/assets/documents/180376%20https://www.fema.gov/media-library/assets/documents/180374.

Gabrielli, Andrea. 2019. "A Neural Network Boosted Double over-Dispersed Poisson Claims Reserving Model." *SSRN Electronic Journal*, April. https://doi.org/10.2139/ssrn.3365517.

Gabrielli, Andrea, Ronald Richman, and Mario V. Wüthrich. 2019. "Neural network embedding of the over-dispersed Poisson reserving model." *Scandinavian Actuarial Journal*. https://doi.org/10.1080/03461238.2019.1633394.

Goldburd, Mark, Anand Khare, Dan Tevet, and Dmitriy Guller. 2020. *Generalized Linear Models for Insurance Rating*. Second Edi. Casualty Actuarial Society. www.casact.org.

Guo, Cheng, and Felix Berkhahn. 2016. "Entity Embeddings of Categorical Variables." *arXiv* arXiv:1604. http://arxiv.org/abs/1604.06737.

Klinker, Fred. 2010. "Generalized Linear Mixed Models for Ratemaking: A Means of Introducing Credibility into a Generalized Linear Model Setting." *Casualty Actuarial Society E-Forum, Winter 2011 Volume 2* 2 (1): 1–25. http://scholar.google.com/scholar?hl=en%7B/&%7DbtnG=Search%7B/&%7Dq=intitle:Generalized+Linear+Mixed+Models+for+Ratemaking+:+A+Means+of+Introducing+Credibility+into+a+Generalized+Linear+Model+Setting%7B/#%7D0.

Kuo, Kevin. 2019. "Deeptriangle: A deep learning approach to loss reserving." *Risks* 7 (3). https://doi.org/10.3390/risks7030097.

———. 2020. "Individual Claims Forecasting with Bayesian Mixture Density Networks," March. http://arxiv.org/abs/2003.02453.

Perla, Francesca, Ronald Richman, Salvatore Scognamiglio, and Mario V. Wüthrich. 2020. "Time-Series Forecasting of Mortality Rates using Deep Learning." *SSRN Electronic Journal*.

Richman, R. 2018. "AI in Actuarial Science." *SSRN Electronic Journal*, October. https://doi.org/10.2139/ssrn.3218082.

Richman, R, and Mario V. Wüthrich. 2019. "A neural network extension of the Lee-Carter model to multiple populations." *Annals of Actuarial Science*. https://doi.org/10.1017/S1748499519000071.

Schelldorfer, Jürg, and Mario V. Wüthrich. 2019. "Nesting Classical Actuarial Models into Neural Networks." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3320525.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention is all you need." In *Advances in Neural Information Processing Systems*, 2017-Decem:5999–6009. http://arxiv.org/abs/1706.03762v5.

Wüthrich, Mario V, and Michael Merz. 2019. "Yes, we CANN!" *ASTIN Bulletin: The Journal of the IAA* 49 (1): 1–3.