Unsupervised learning for efficient underwriting

Per Wilhelmsson
Pricing actuary

Elena Dalla Torre
Master's student

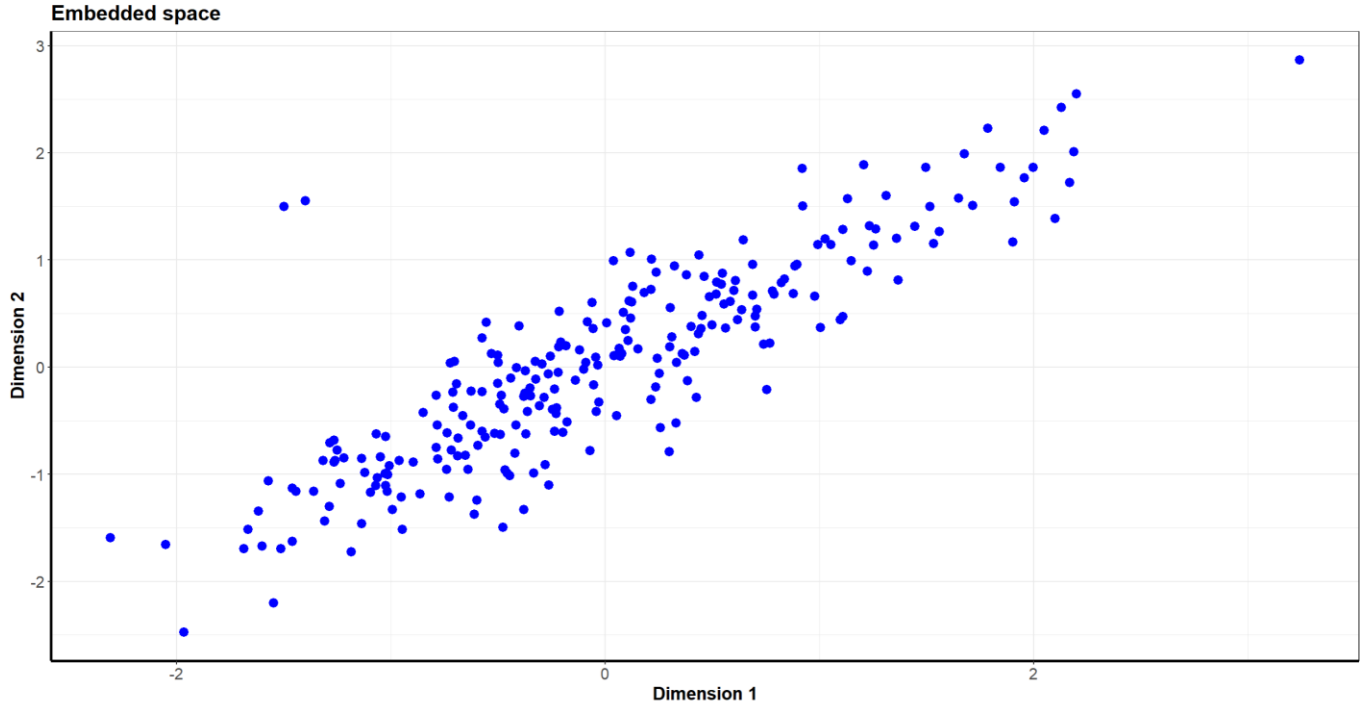
Länsförsäkringar

Agenda

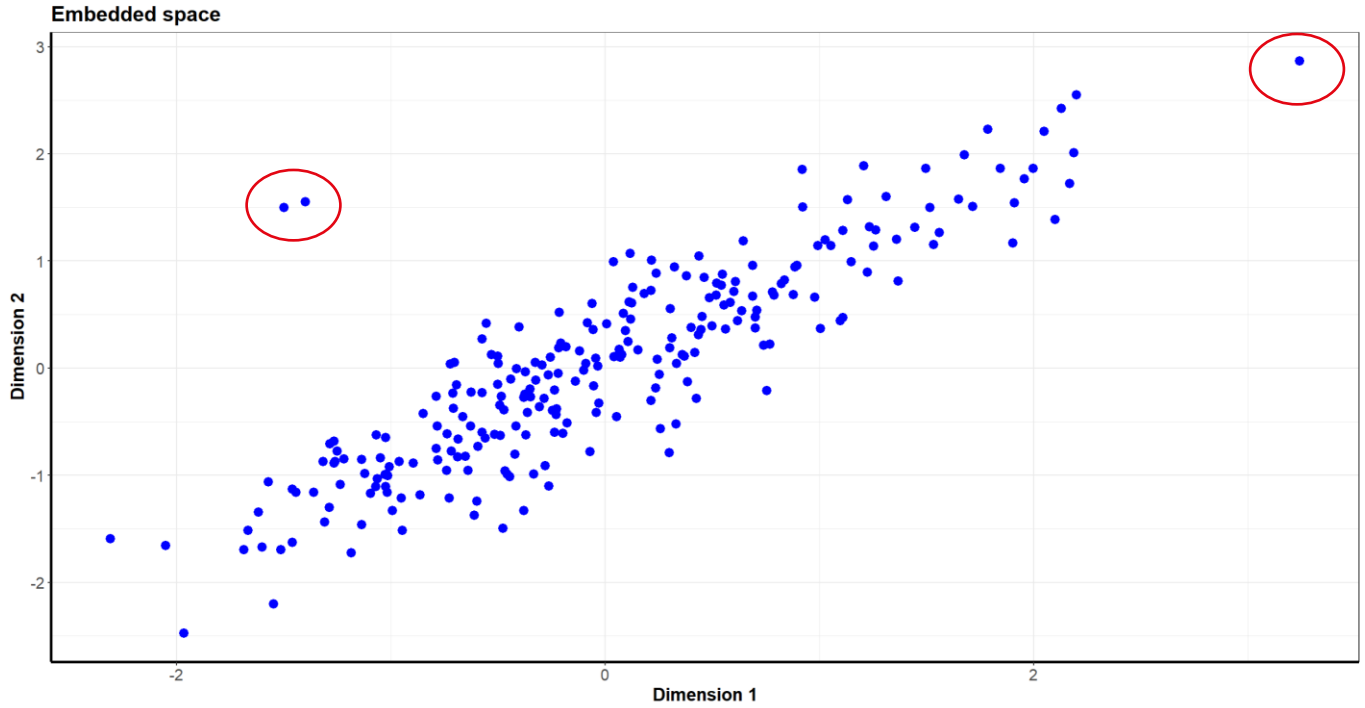
- **Embed an insurance, why?**
  - *Find unique objects*
  - *Find similar objects*
  - *Create pricing features*
  - *Measure information density*

- **How to embed insurances and create a uniqueness value.**
  - *Embedding - PCA and autoencoders*
  - *Create a uniqueness score - PCA and autoencoders*

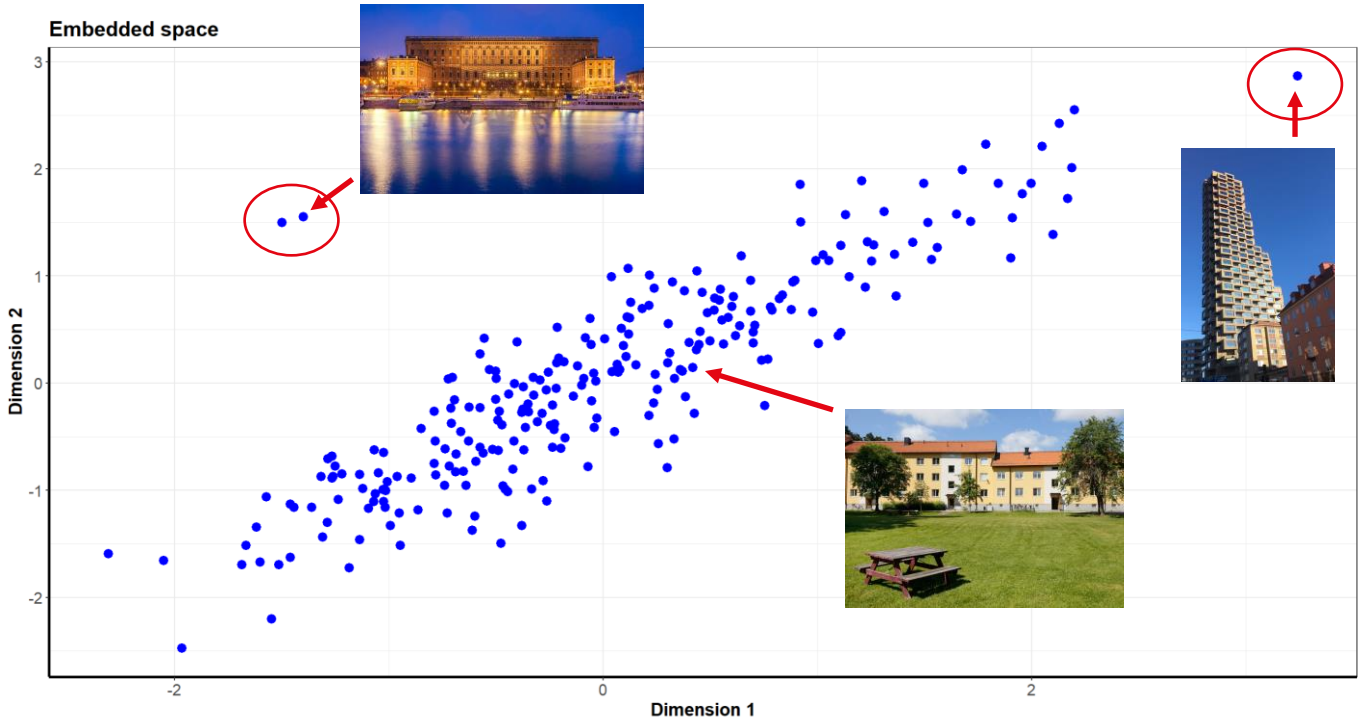
Embeddings: A vectorized representation of the original data



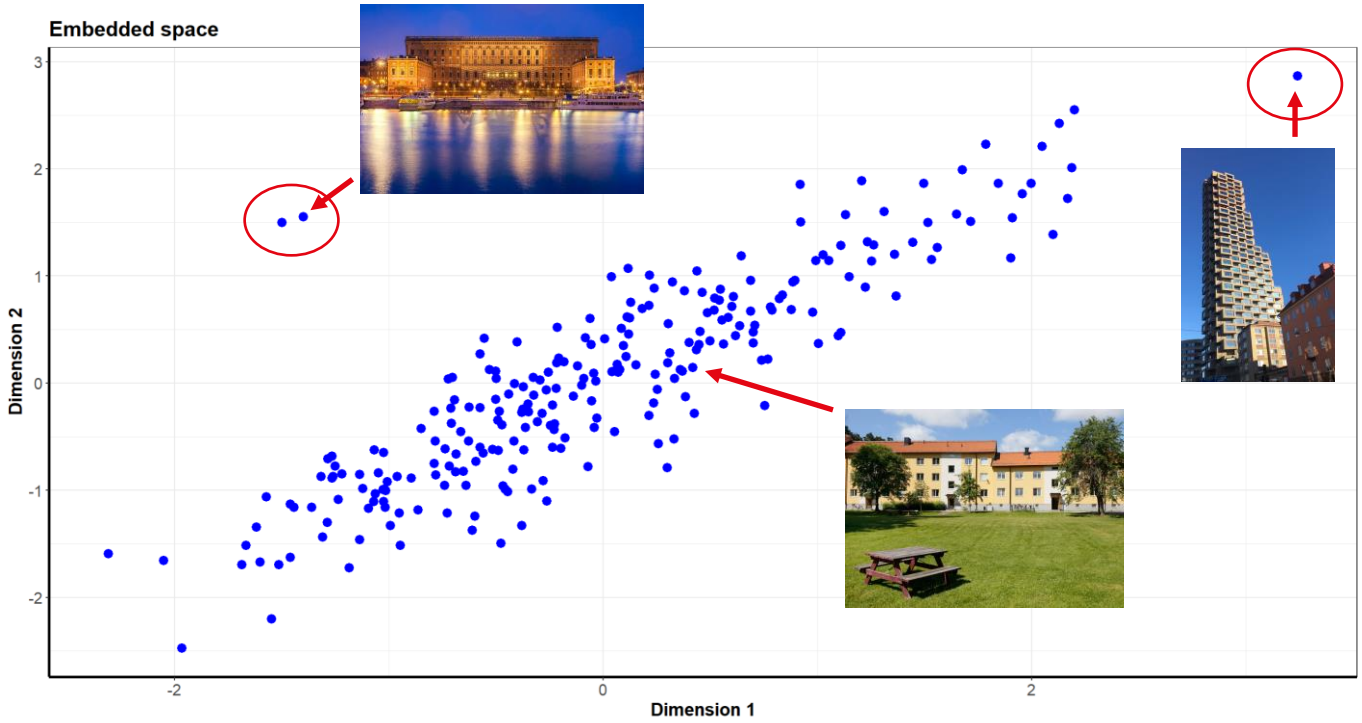
1. Find unique objects



1. Find unique objects



1. Find unique objects

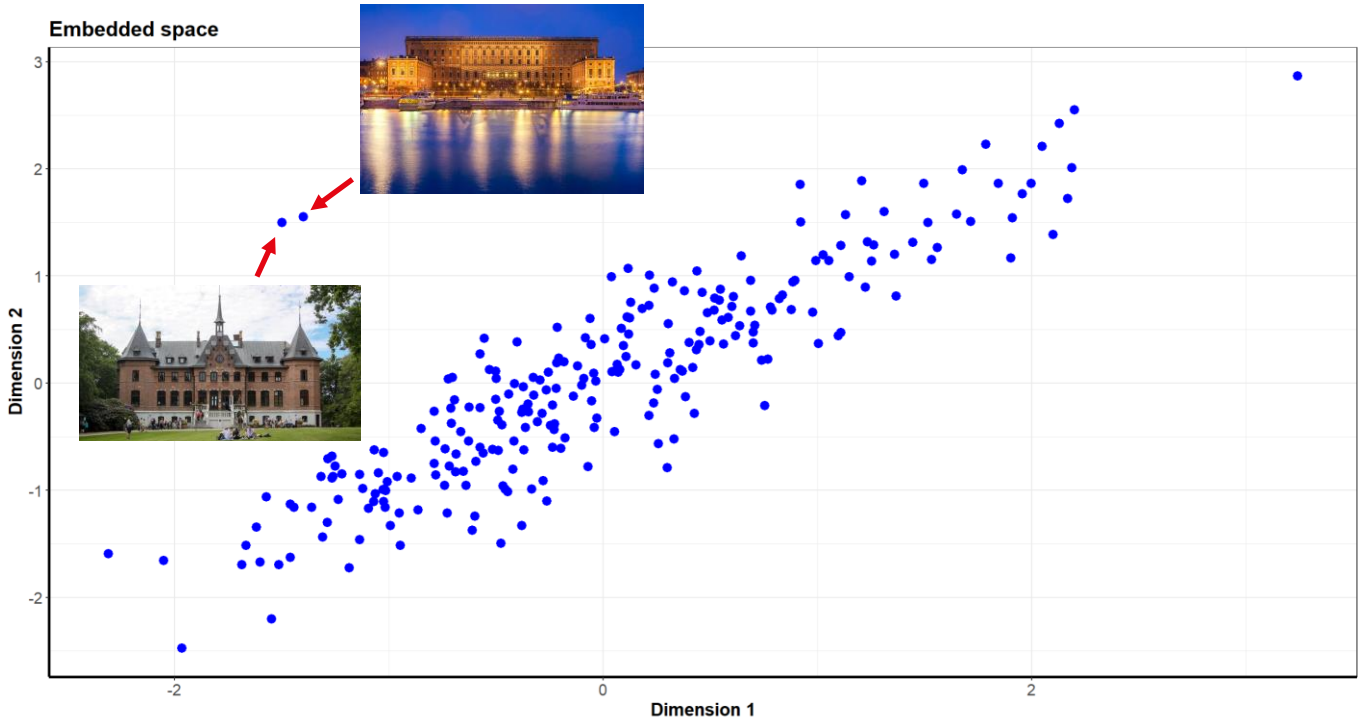


Can uniqueness be a pricing feature in and of itself?

Unique objects should be manually underwritten to a larger extend than a non-unique object.

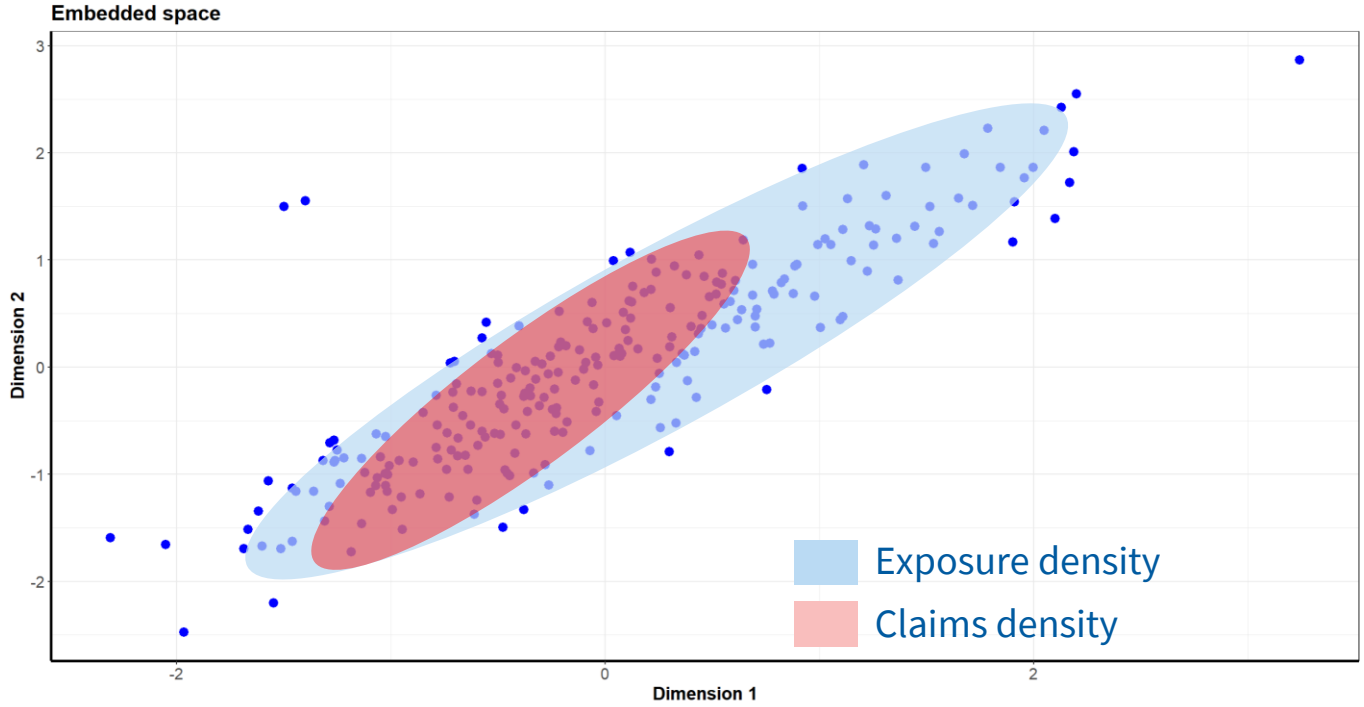
Find errors in the insurance registration.

2. Find the most similar objects to an object



When manually underwriting an insurance, it is helpful to see examples of similar insurances

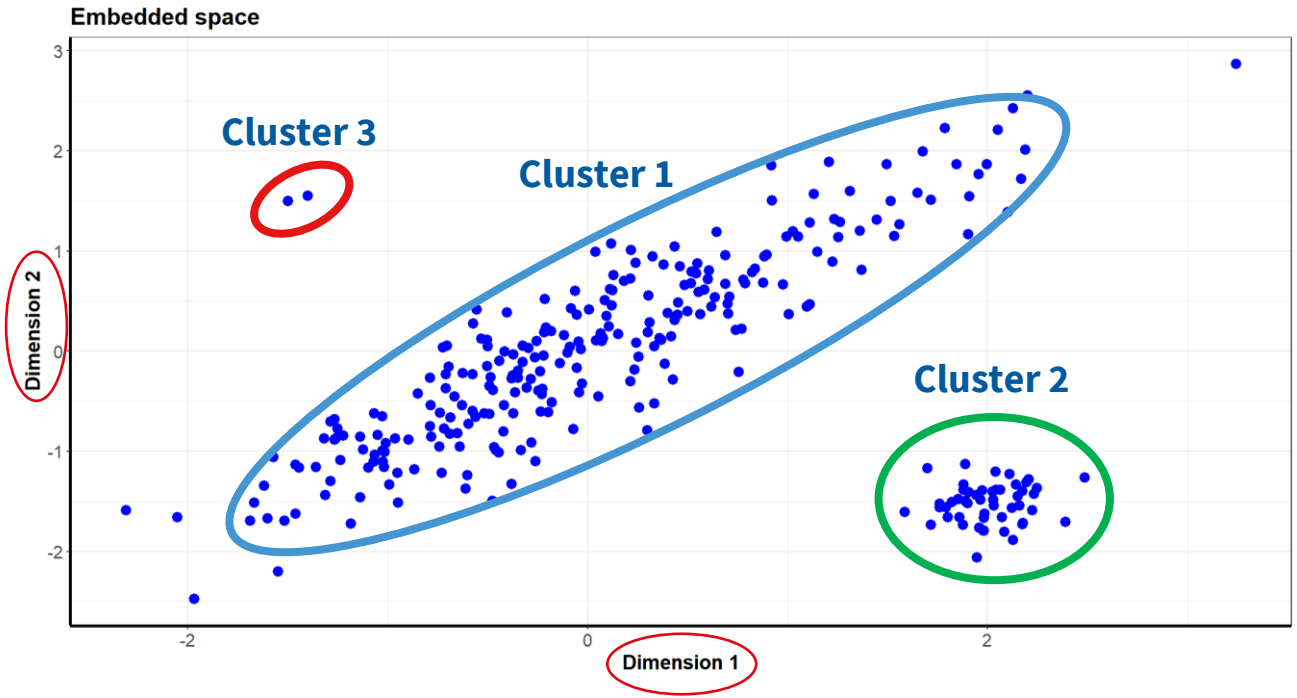
3. Combine exposure and claims densities to measure pricing uncertainty



Areas with a high claims-to-exposure ratio are areas where pricing algorithms are stable.

Areas with a low claims-to-exposure ratio are areas where there is high uncertainty about what a risk correct premium is.

4. The vector embeddings or clusters can be used as pricing features



In addition to use uniqueness as a pricing feature, we can use the vector embeddings or clusters as pricing features, extending the power of GLMs to offer fair premiums.

Business applications:

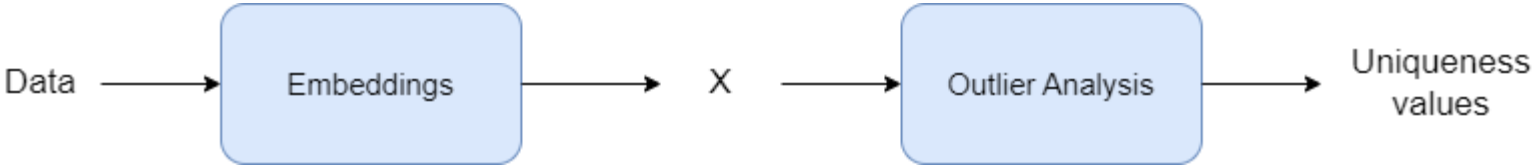
1. **Find unique objects, govern UW mandates and use as a pricing feature. Alert salespeople when anomalies are registered in insurances.**

2. **Find similar objects and use as reference.**

3. **Measure how certain our pricing algorithms can be on a given object.**

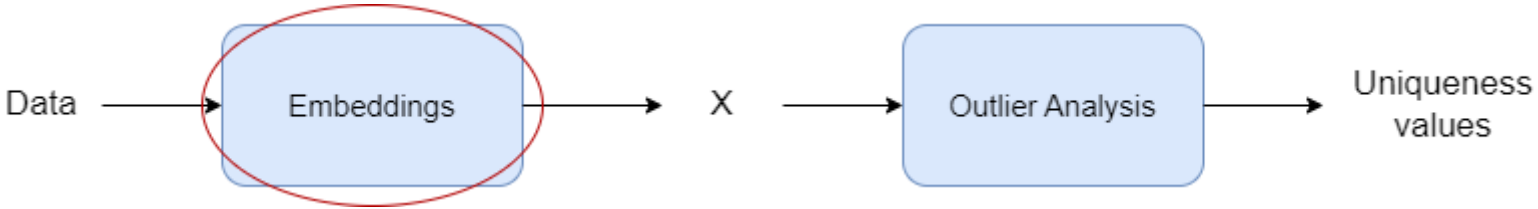
4. **Use embeddings and/or clusters as pricing features**

How to embed insurances and create a uniqueness value



- Embed insurances as outlier detection methods work with numerical data → find a numerical representation of the categorical variables in the data.
- Outlier detection methods output outlier scores to be used as uniqueness values.

Numerical Representation of Categorical Variables - Embedding

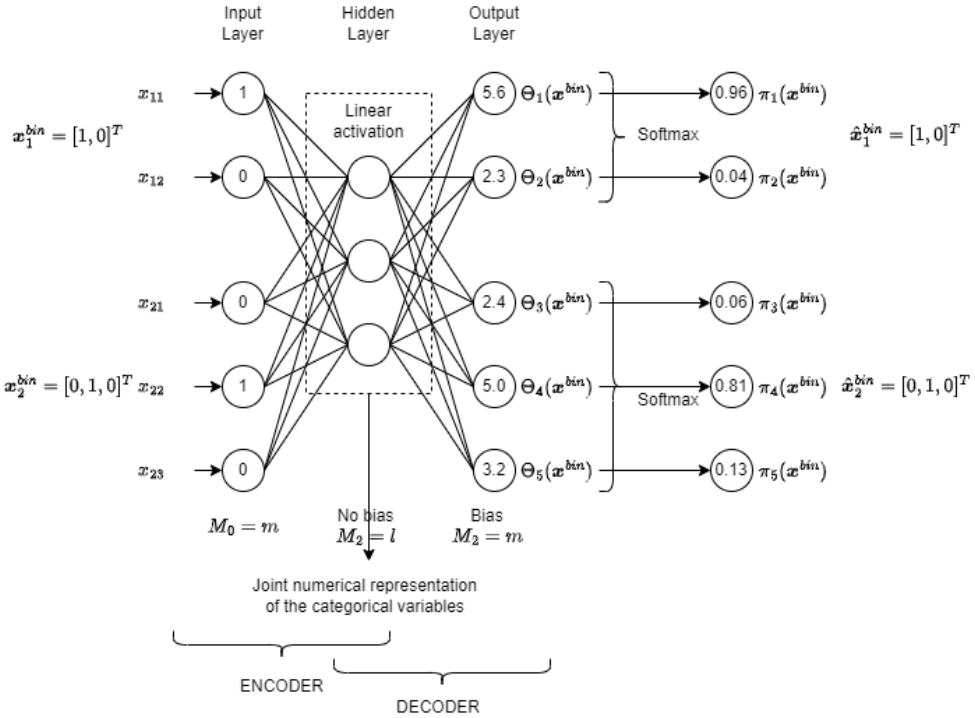


- PCAmix (linear autoencoder)
- Non-linear autoencoder

Principal Component Analysis of Mixed Data (PCAmix)

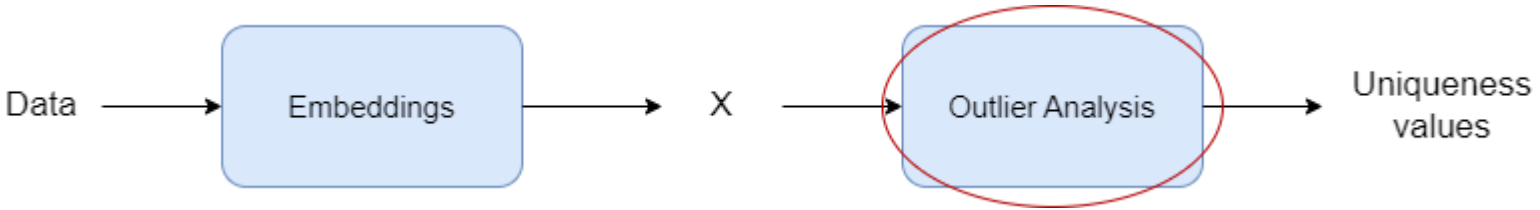
- Principal component method: finds numerical representation of mixed data set of both numerical and categorical variables with maximum link with the original data set.
- PCAmix can be thought of as a linear autoencoder: all activation functions are linear, mean squared error loss used for training.

Non-linear Autoencoder



- Linearity assumption is lifted using *softmax* activation in the output layer.
- Autoencoder takes as input the binarized categorical variables (one-hot encoding).
- Autoencoder outputs the probabilities of the input observation to be in each category of the categorical variables.
- Softmax is applied to groups of neurons corresponding to the categories of each categorical variable.
- Output of encoder layer is used as a joint numerical representation of the categorical variables of dimensionality l.

Outlier Analysis

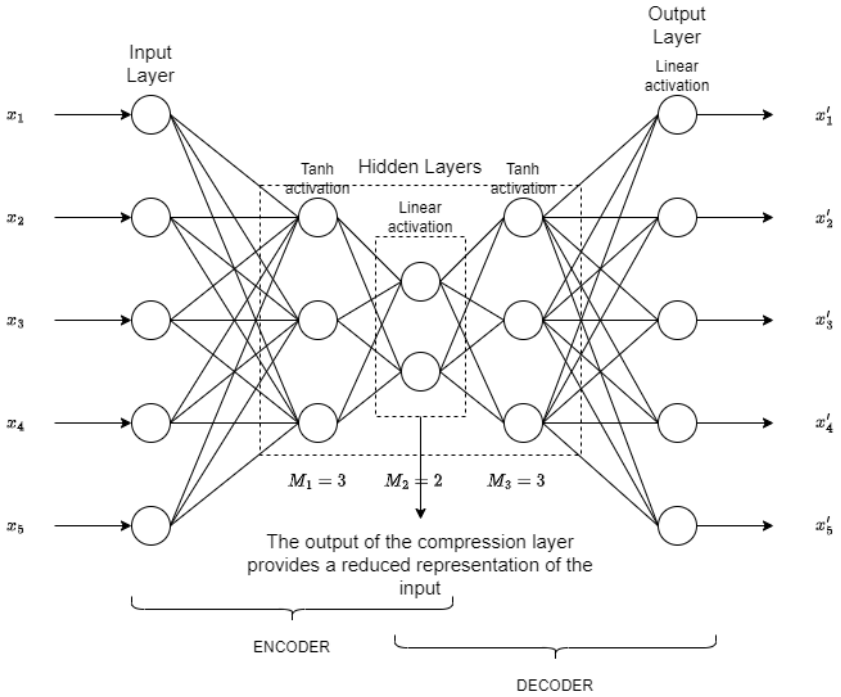


- PCA (linear autoencoder)
- Non-linear autoencoder

Principal Component Analysis (PCA)

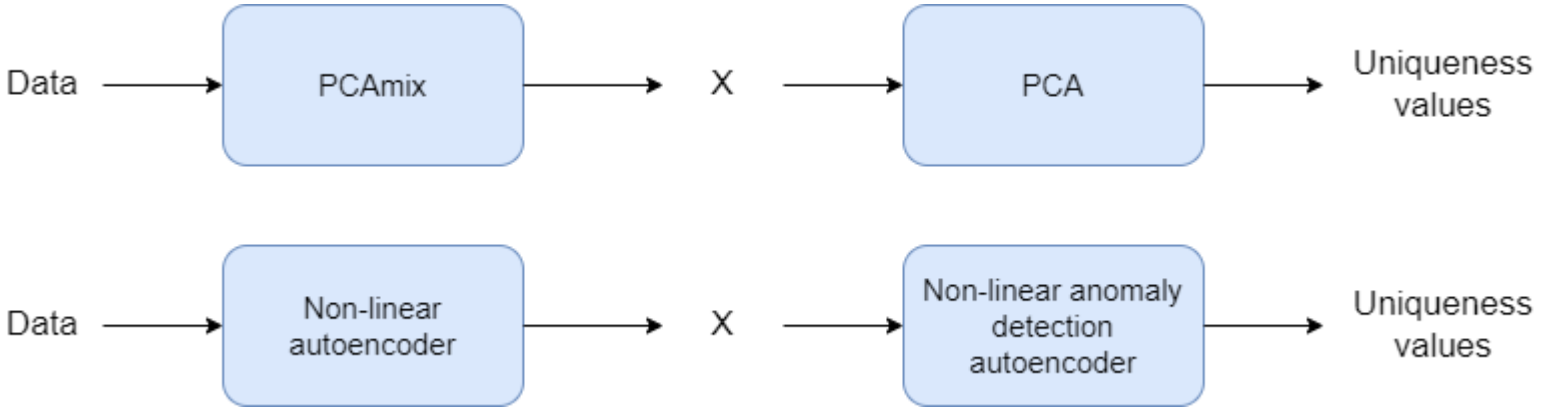
- Outlier score of an observation is the sum of its squared distances from the mean along each eigenvector, each divided by the corresponding eigenvalue.
- Soft approach to PCA: all eigenvectors are used.
- Equivalent to scores found using the Mahalanobis distance.

Non-linear Anomaly Detection Autoencoder



- Outlier scores are computed as reconstruction errors between the input and the output.
- Tendency to overfit → important to use suitable regularization techniques (dropout, pretraining).

Linear Method vs Non-linear Method



Analysis of the Results

- The linear method tends to identify as unique, observations belonging to rare categories of categorical variables or with extreme values for the numerical variables.
- The non-linear method identifies observations that are unique because of less obvious combinations of variables.
- The superior performance of the non-linear method is explained by its ability to capture non-linear dependencies among the variables.