ANNUAL MEETING

November 9-12, 2020 • Online Event

Determining Vehicle Symbols Using Machine Learning Techniques

Giorgio A. Spedicato & Marco De Virgilis

Presenters

Giorgio Alfredo Spedicato, Ph.D FCAS FSA CSPA C.Stat
Data Science Manager,
Unipol Group

Marco De Virgilis,
Senior Actuarial Data Scientist,
The Allstate Corporation

Disclaimer

The views and opinions expressed in this presentation are those of the authors’ and do not necessarily reflect the position of the organizations of which they are part.

Agenda

- Vehicle Symbols Explanation
- Analysis and Methodology
- Algorithms Implemented
- Application
- Model Comparisons
- Conclusion

Vehicle Symbols

Vehicle Symbols

- Vehicle Symbols (VS) are codes that **group** vehicles experiencing **similar** loss costs. In practice, a code is assigned to a vehicle **which corresponds to a loss relativity.** The VS assigned to a given vehicle type may also vary by peril.

- Insurers writing motor perils coverage would typically charge vehicles belonging to the **same VS** group the **same price** — all policyholder characteristics being equal.

- A company may develop VS by itself or use those provided by Rating Bureaus.

Vehicle Symbols

- Determining VS is an important task in developing a sound ratemaking framework in **motor insurance**.

- Recent improvements have paved the way for more sophisticated algorithms that make more extensive use of data, reaching unprecedented levels of **performance**.

- Our aim is to show how a VS estimation exercise is carried out by exploiting unsupervised and supervised **Machine Learning** methods.

Analysis and Methodology

Analysis and Methodology

- Explore several ML techniques to either **directly group** vehicles into groups or to uncover **latent dimensions** that summarize their essential characteristics.

- Compare the different methodologies **quantitatively**, in terms of predictive performance and **qualitatively**, in terms of practicality and communicability.

- The **Gini Index** will quantify the predictive performance, while one-way plots will depict the relation between the clusters (or the latent dimensions) and the insurance risk.

Algorithms Implemented

Algorithms Implemented

Dimensionality Reduction Algorithms

- Principal Component Analysis
- t-Distributed Stochastic Neighbor Embedding
- Generalized Low Rank Models
- Autoencoders

Unsupervised Clustering Algorithms

- K-Means
- Combined Multivariate Gaussian-Multinomial Mixture Models
- Ordering Points To Identify the Clustering Structure

Supervised Clustering Algorithms

- Tree-based Approaches

Algorithms Implemented

- **Clustering**
  - K-Means
  - Combined Multivariate Gaussian-Multinomial Mixture Models
  - Ordering Points To Identify the Clustering Structure
  - Tree-based Approaches
- **Dimensionality Reduction**
  - Principal Component Analysis
  - t-Distributed Stochastic Neighbor Embedding
  - Generalized Low Rank Models
  - Autoencoders

Dimensionality Reduction

Principal Component Analysis (PCA)

- Principal Component Analysis (PCA) is an approach for deriving a **low-dimensional** set of features from a large set of variables.
- PCA finds a small number of dimensions that keeps the initial dataset **variation** by computing a **linear** combination of the initial features.
- These linear combinations are called *Principal Components*.
- $Z_k = \phi_{1k}X_1 + \phi_{2k}X_2 + \cdots + \phi_{ik}X_i$

t-Distributed Stochastic Neighbor Embedding (t-SNE)

- t-SNE is a **non-linear** technique, while, PCA applies a linear transformation to the original data.

- Another important distinction is that, whereas, PCA tries to preserve the global similarities, t-SNE is more concerned with preserving **local** similarities.

- The algorithm optimizes a cost function that computes the **Euclidean distance** between the high-dimensional points and the low-dimensional points.

Generalized Low Rank Models (GLRM)

- GRLMs are a matrix factorization technique that represents a dimension reduction able to handle **mixed**-data matrices.

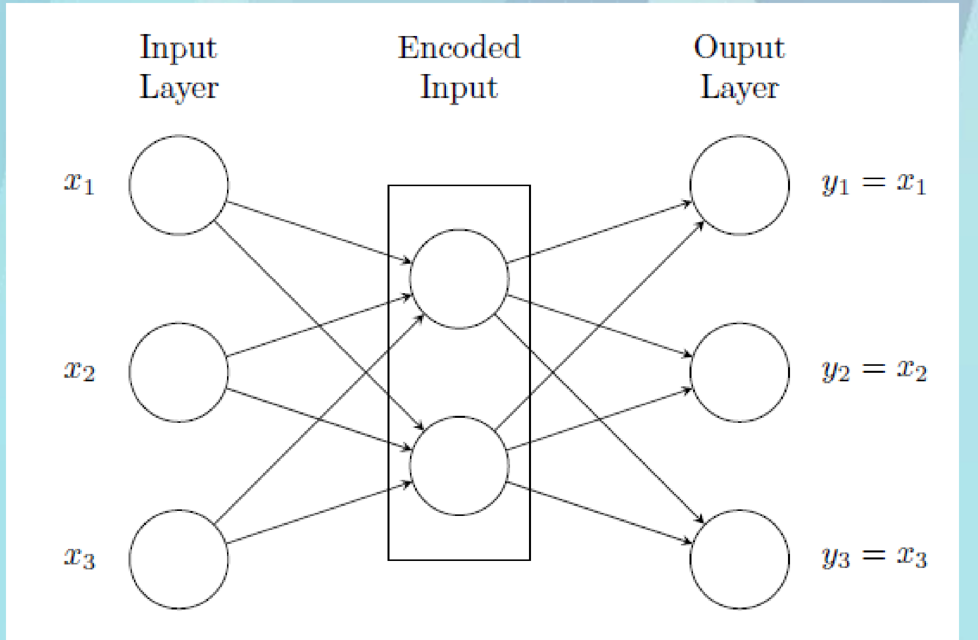
- GLRMs are commonly used as extension of the PCA technique, to naturally handle **mixed** data sets containing ordinal, categorical, Poisson and Boolean data types.

- They approximate an input data matrix $X_{m,n}$ by **projecting** it in a reduced low rank form.

- $X_{m,n} \approx A_{m,k} * Y_{k,n}$

Autoencoders

- Autoencoders are a type of Artificial Neural Networks used to learn feature representations in an unsupervised manner
- They can be though as very powerful **non-linear** generalization of PCA.



Unsupervised Clustering

K-Means

- The K-Means algorithm is a clustering method which aims to **partition** a set of data points into *k* clusters, in which each observation belongs to the clusters with the **nearest** mean.

- It is an **iterative** algorithm that finds clusters by minimizing the **Euclidean distances** between points, hence it minimize the within-cluster variances.

- Extensions of the algorithm, namely K-Mods and K-Prototypes can also handle **categorical** variables.

Combined Multivariate Gaussian-Multinomial Mixture Models (Mixmod)

- Mixture models assume that the data are an i.i.d. samples from some population described by a **probability density** function.

- This density function is a finite **mixture** of parametric component density functions (e.g. multinomial or gaussian) where each component models one of the cluster.

- The advantage of using mixture models is that it allows to analyze **all the data** possibilities, **numerical or categorical**, in a unified modeling approach.

Ordering Points To Identify the Clustering Structure (OPTICS)

- Many clustering algorithms, e.g. K-Means, require the input of **series of parameters** in order to identify the clustering structure.
- Density-based approaches overcome this drawback and usually require **less parameters** to identify clusters.
- The OPTICS algorithm makes the all process seemingly **parameter-less**.
- The aim is to either assign each data point to a **cluster** or classify it as **noise**.

Supervised Methods

Tree-based Approaches (CART)

- Classification and Regression Trees (CART) **recursively split** the dataset into smaller subsets that are defined in terms of intervals of the target variable.

- The algorithms are able to unravel interactions between variables and represent them in terms of **hierarchical dependency** structures.

Application

Application

- The research will focus on how **vehicle characteristics** significantly affects the insurance risk, keeping them as the primary point of view.

- This means that each analysis shall lead to a finite number of groups of vehicle that share **similar** characteristics.

- These groups will be analyzed in terms of **claim frequency,** on a dataset used in a Kaggle competition sponsored by Allstate in 2011.

Model Comparisons

Model Comparisons

- The following slides will show the dataset grouped according to the found **VS clusters** and the insurance risk (**claim frequency**).

- We are looking for well-defined and **well-separated cluster** or **monotone relations** between the identified latent variables and the claim frequency.

- A quantitative ranking will be performed evaluating the **Gini Index** as measure of the VS clustering performance.

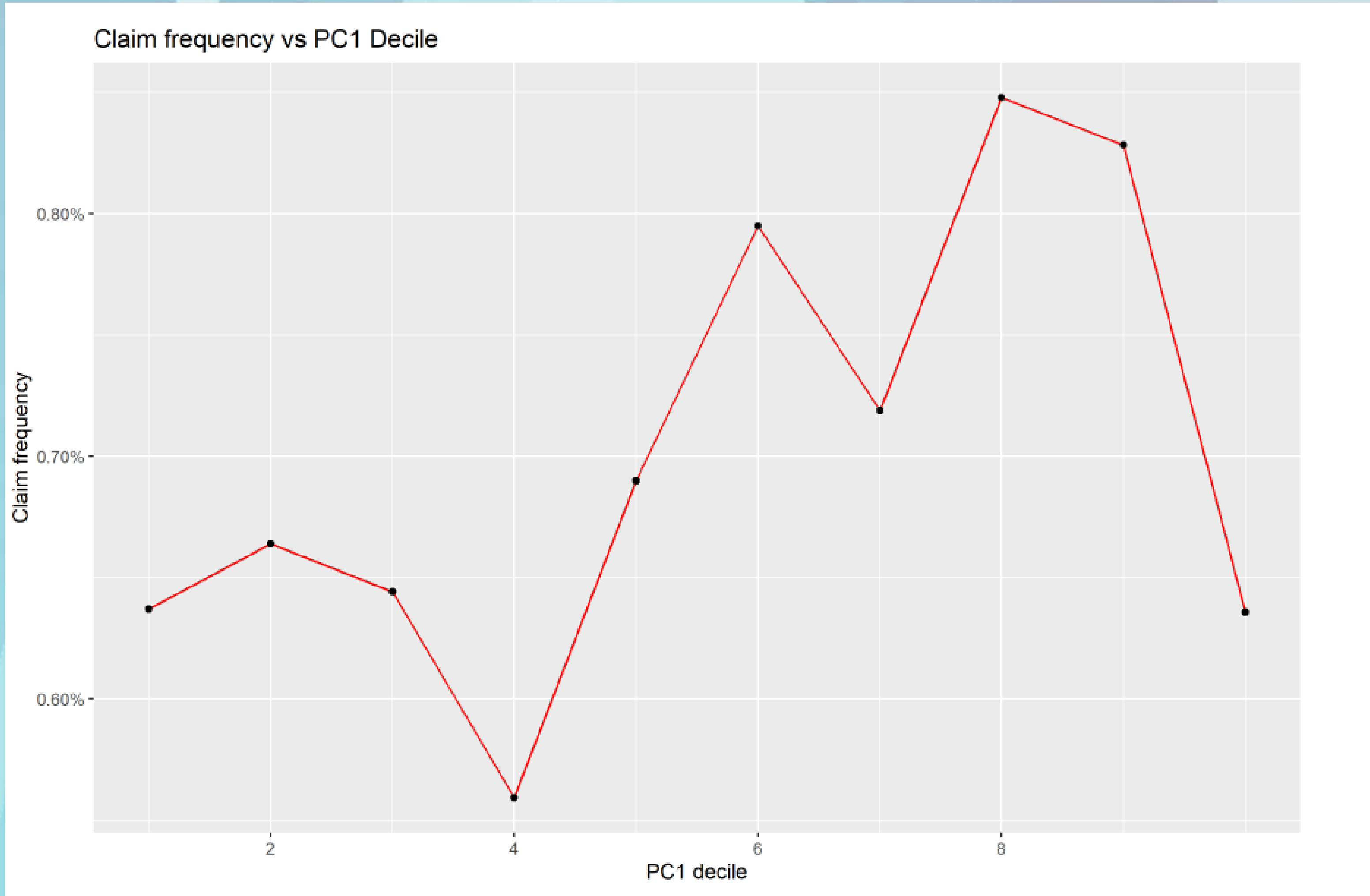
Model Comparisons

- The score given to each VS grouping is based on the quality of the prediction of a **frequency GLM:**
$$E(\lambda_i) = f(x_i) + offset(\ln(Exposure)).$$

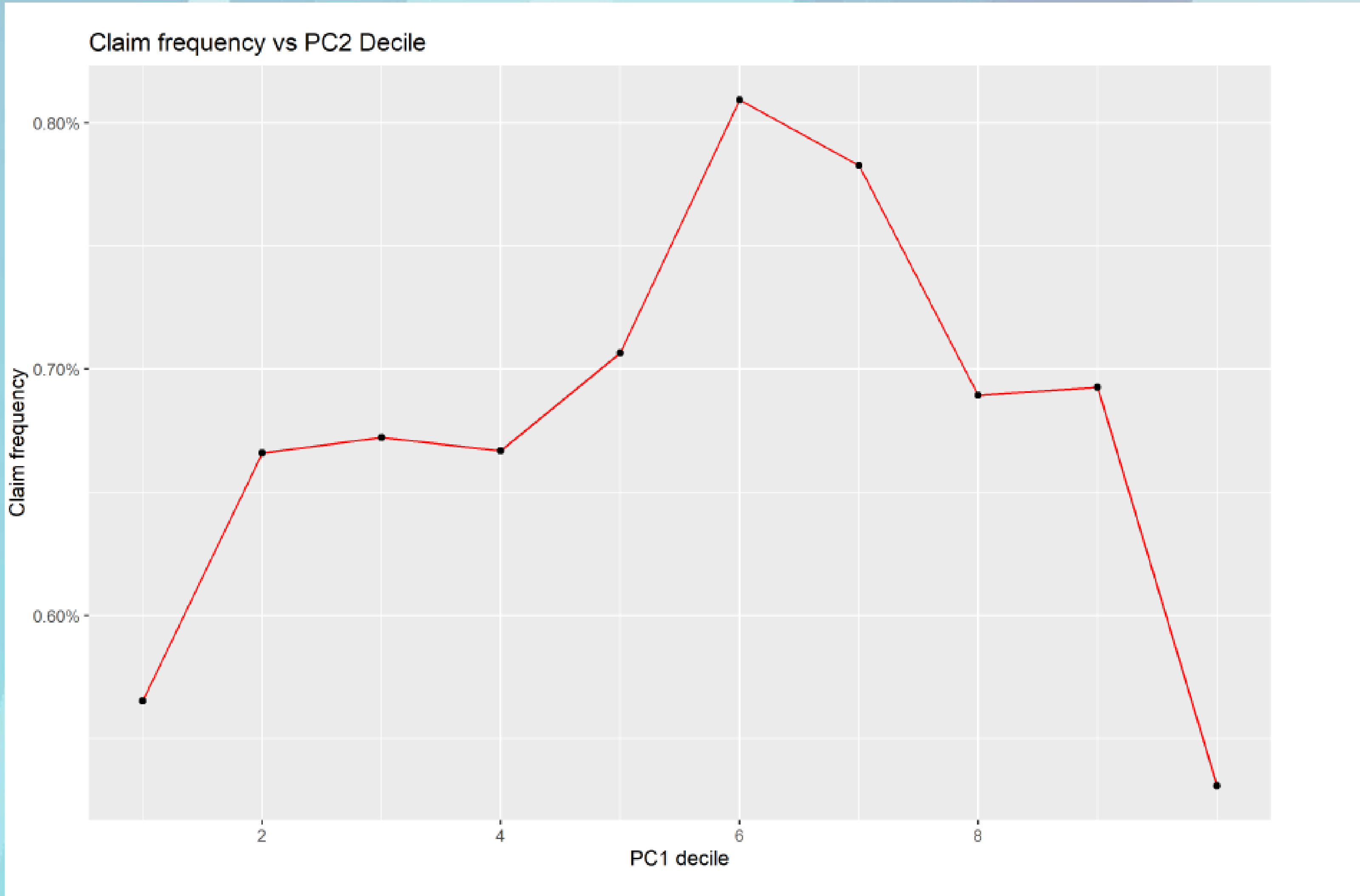
- When the VS model provides k categorical cluster indexes, $f(x_i)$ is the **dummy coefficient** given to each cluster.

- When the model provides k latent dimensions (η_i^k), for **each latent dimension,** we compute a separate GLM, where $f(x_i) = \beta_k * \eta_i^k$. We are assessing whether the *k-th* latent dimension shows a **monotone** relation with the insurance risk and estimate the model performance taking the **highest** Gini index.

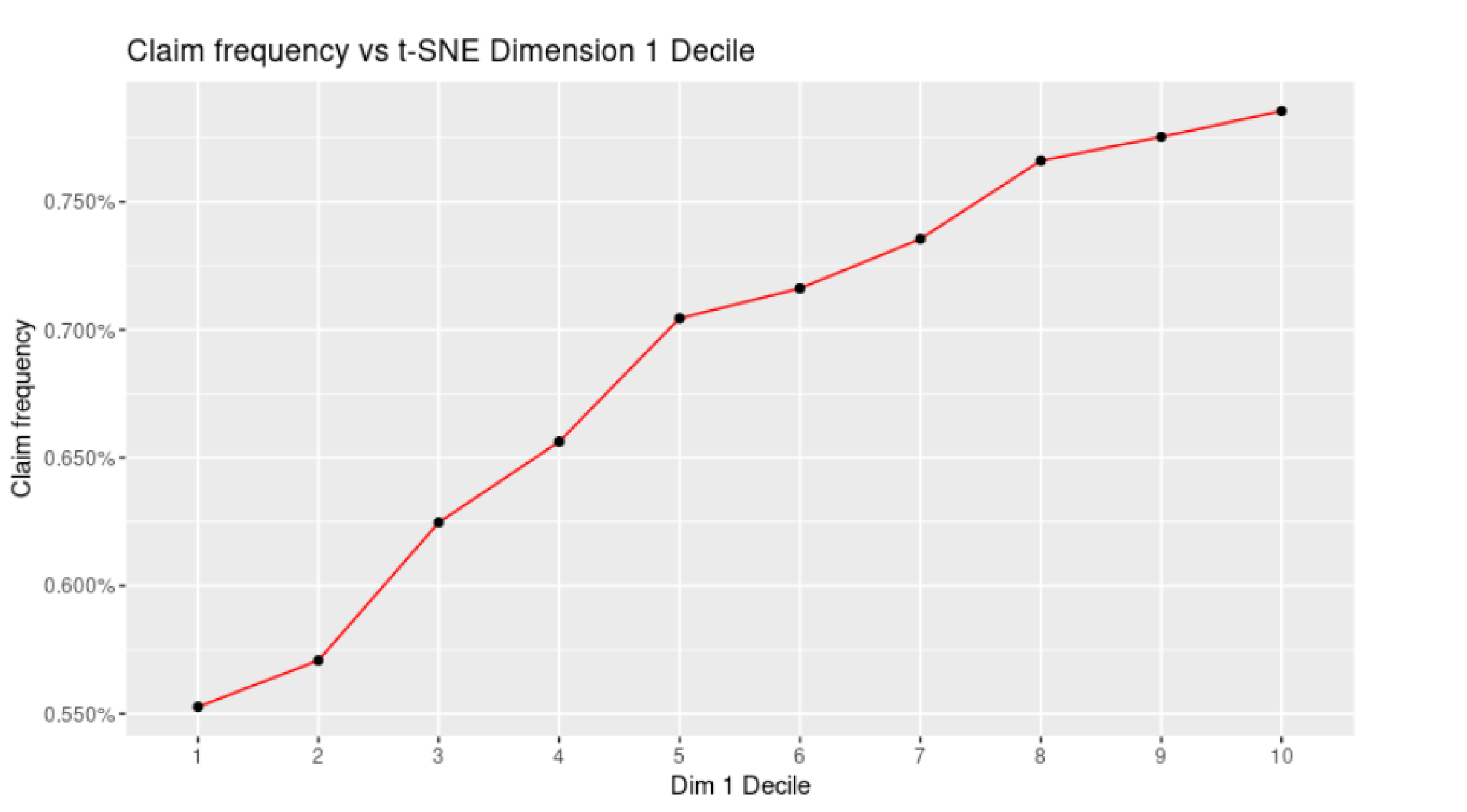
Application of PCA



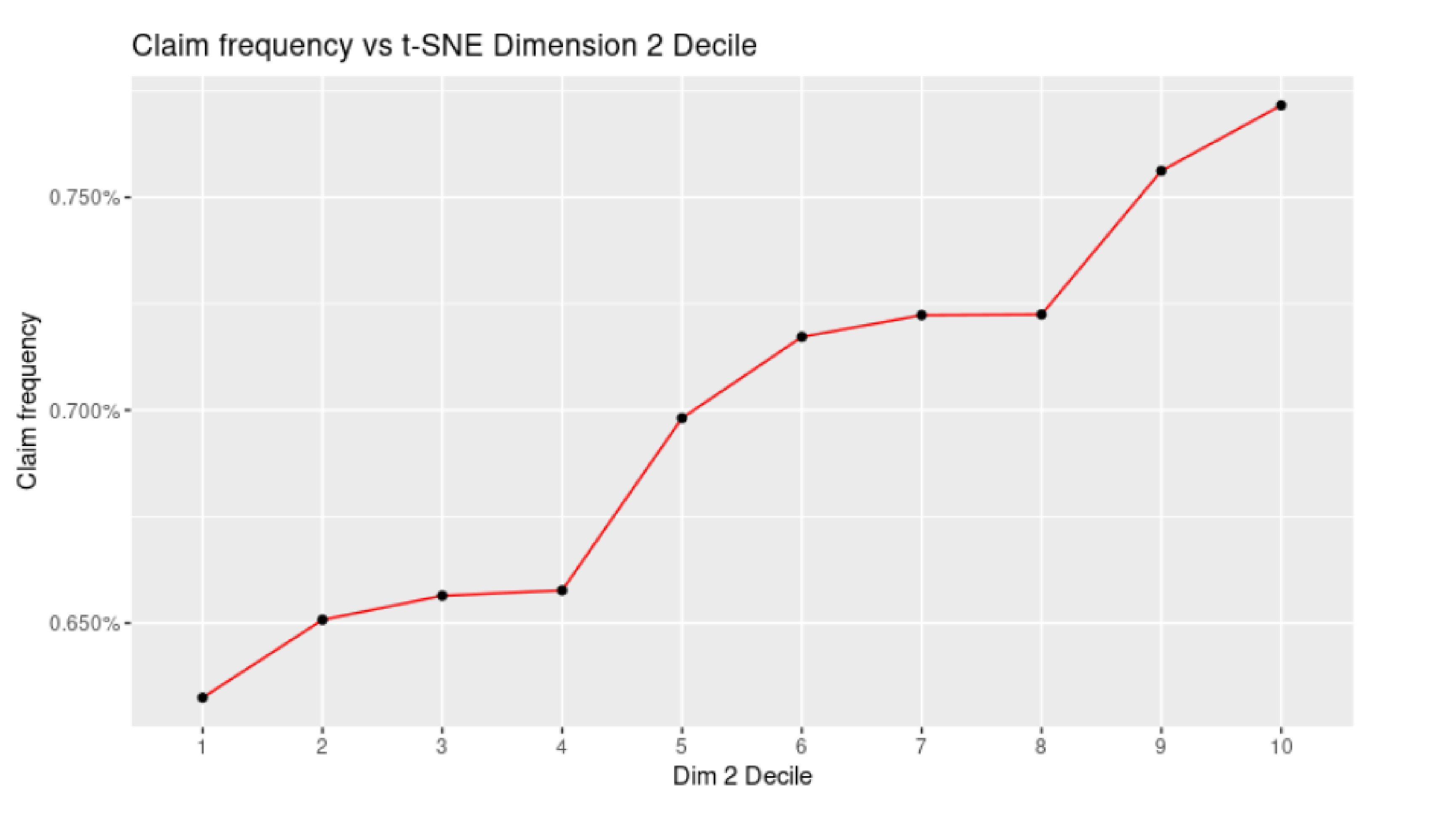
Application of PCA



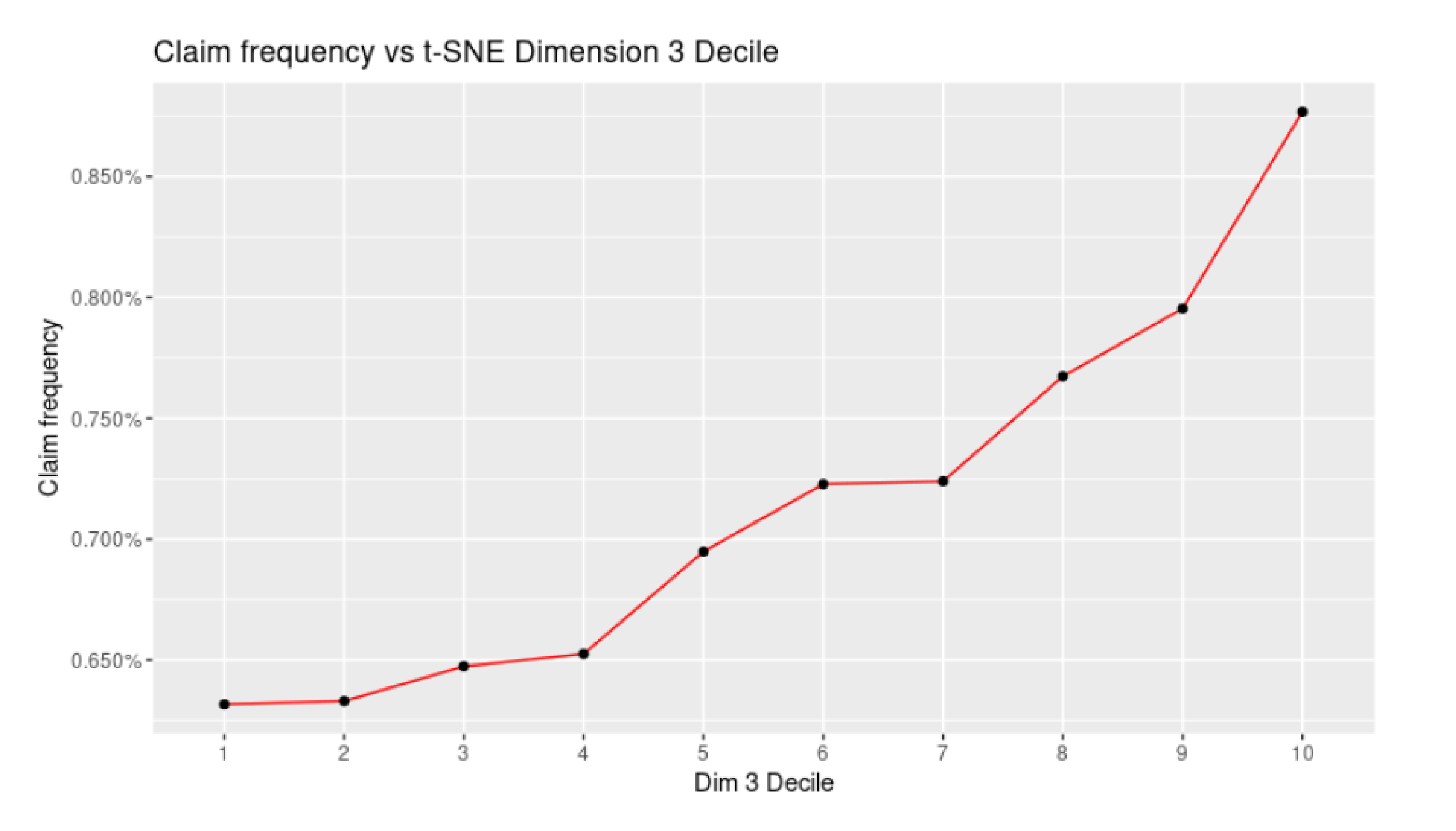
Application of t-SNE



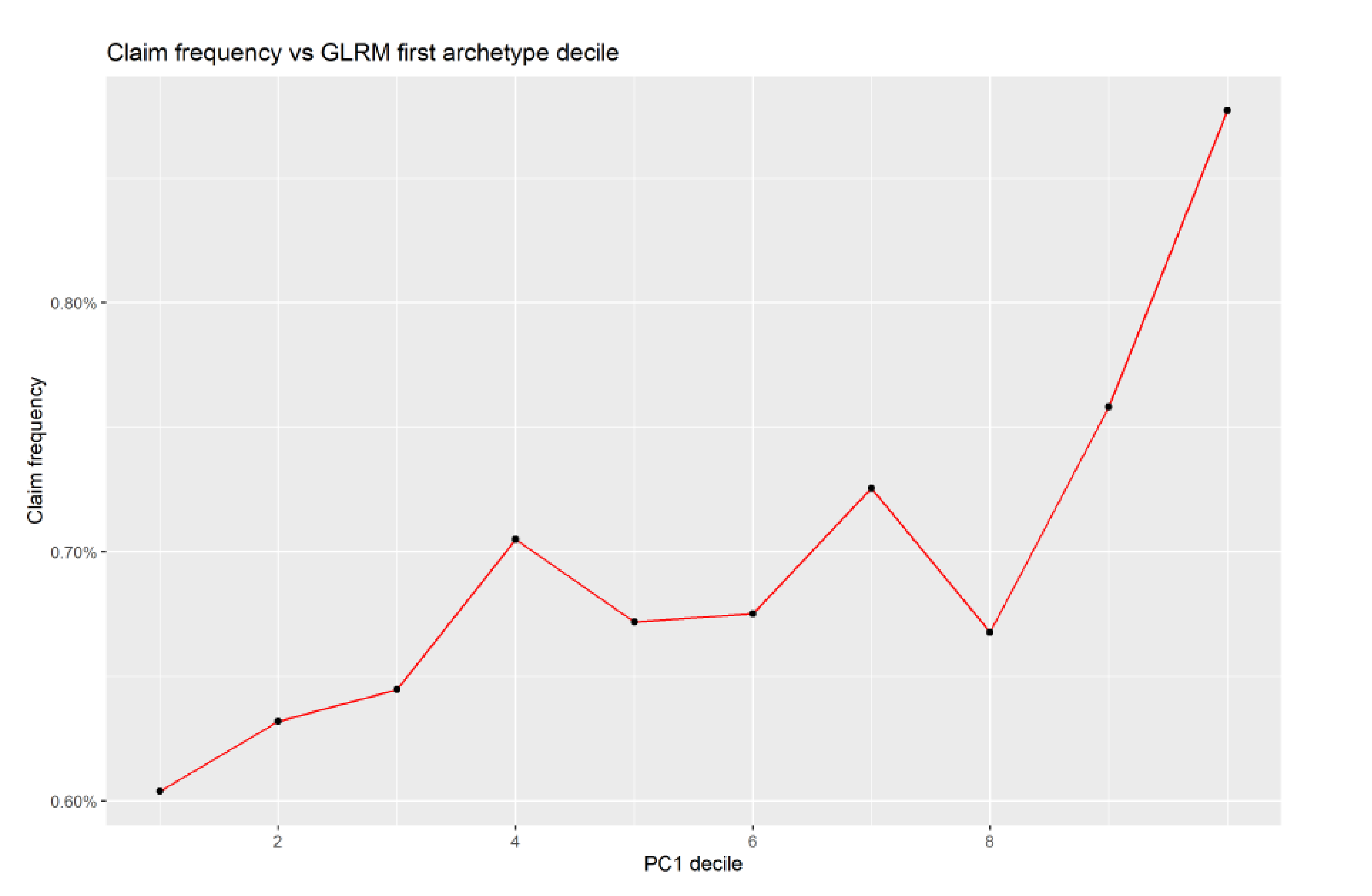
Application of t-SNE



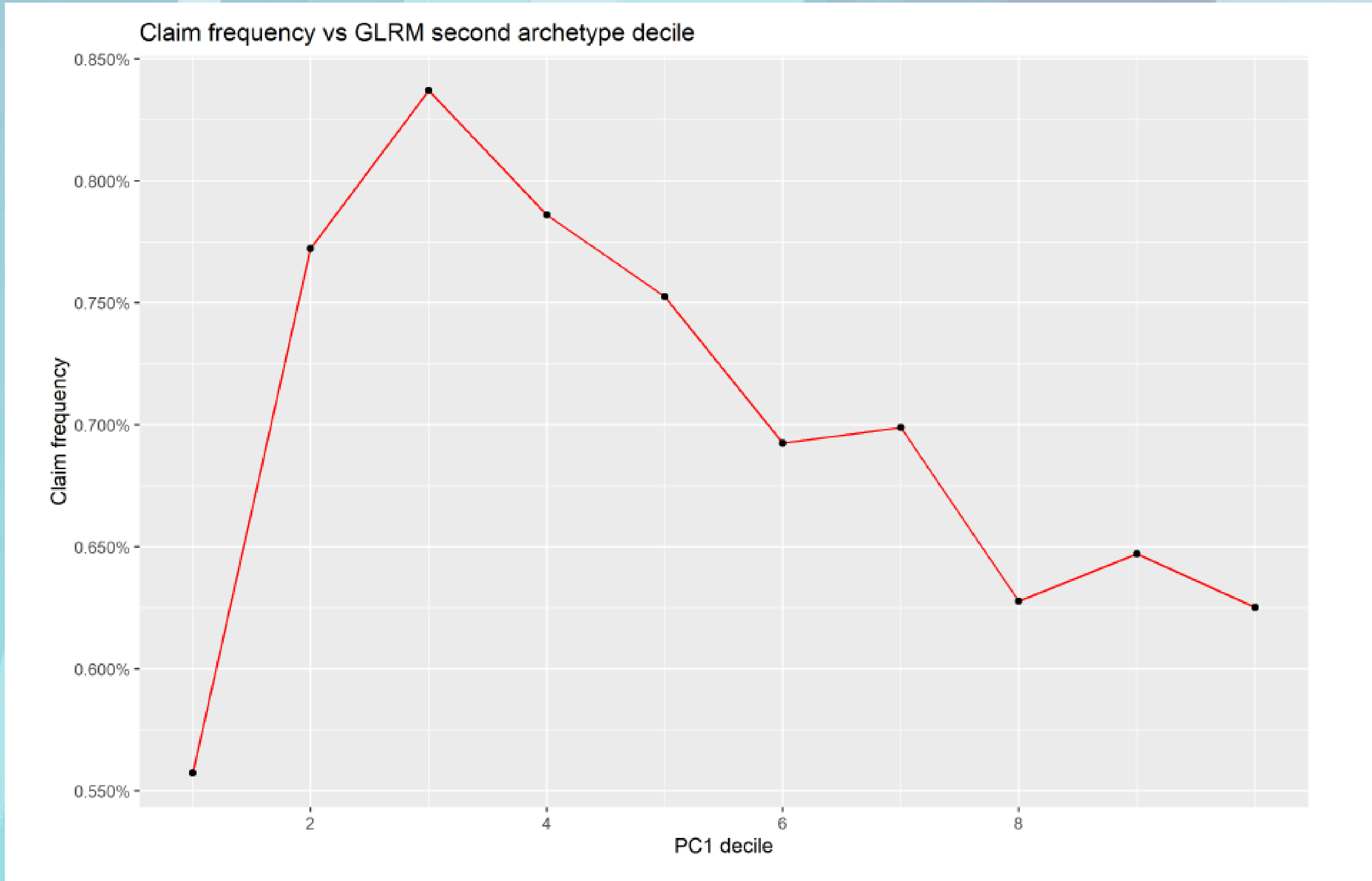
Application of t-SNE



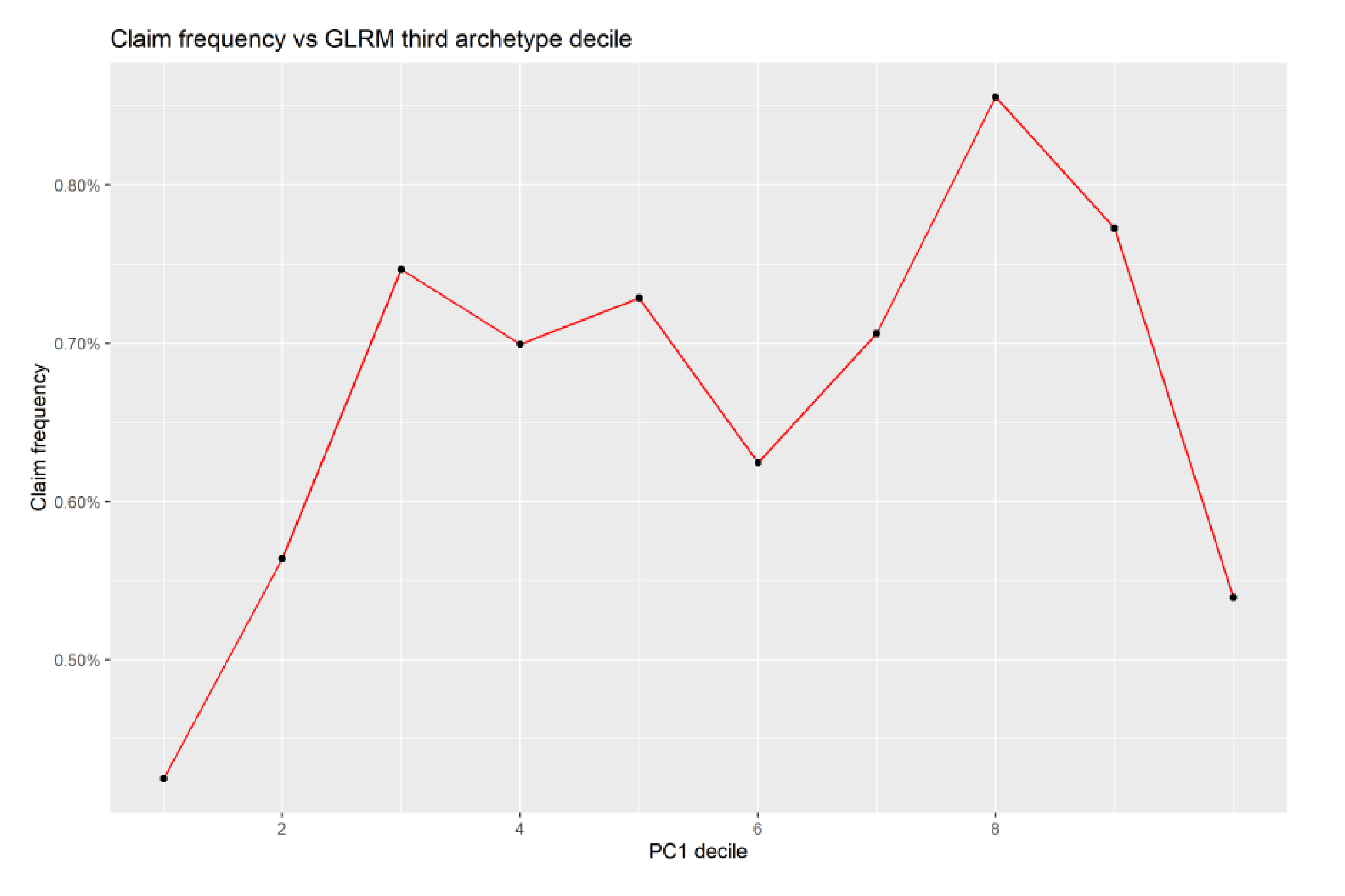
Application of GLRM



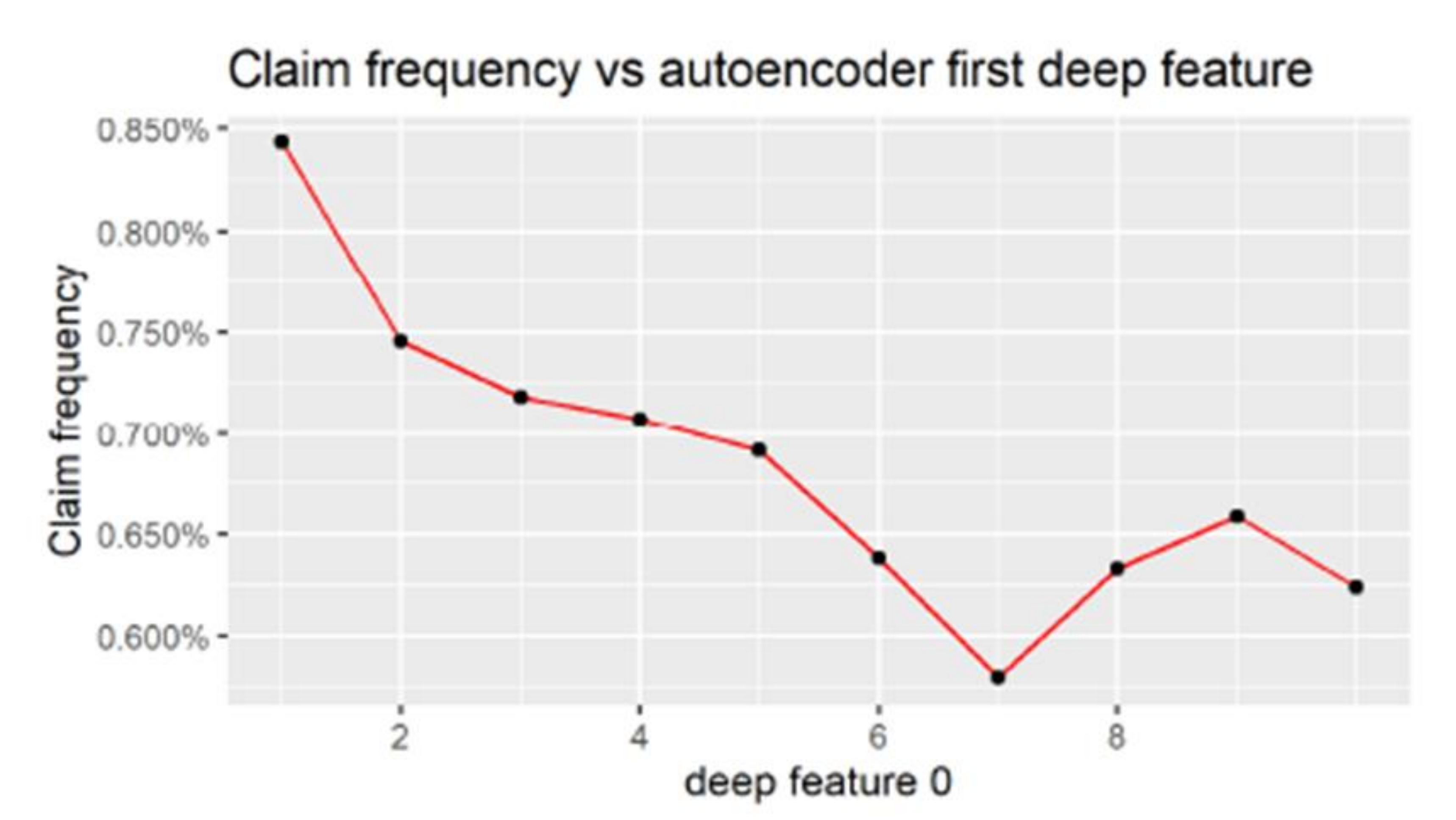
Application of GLRM



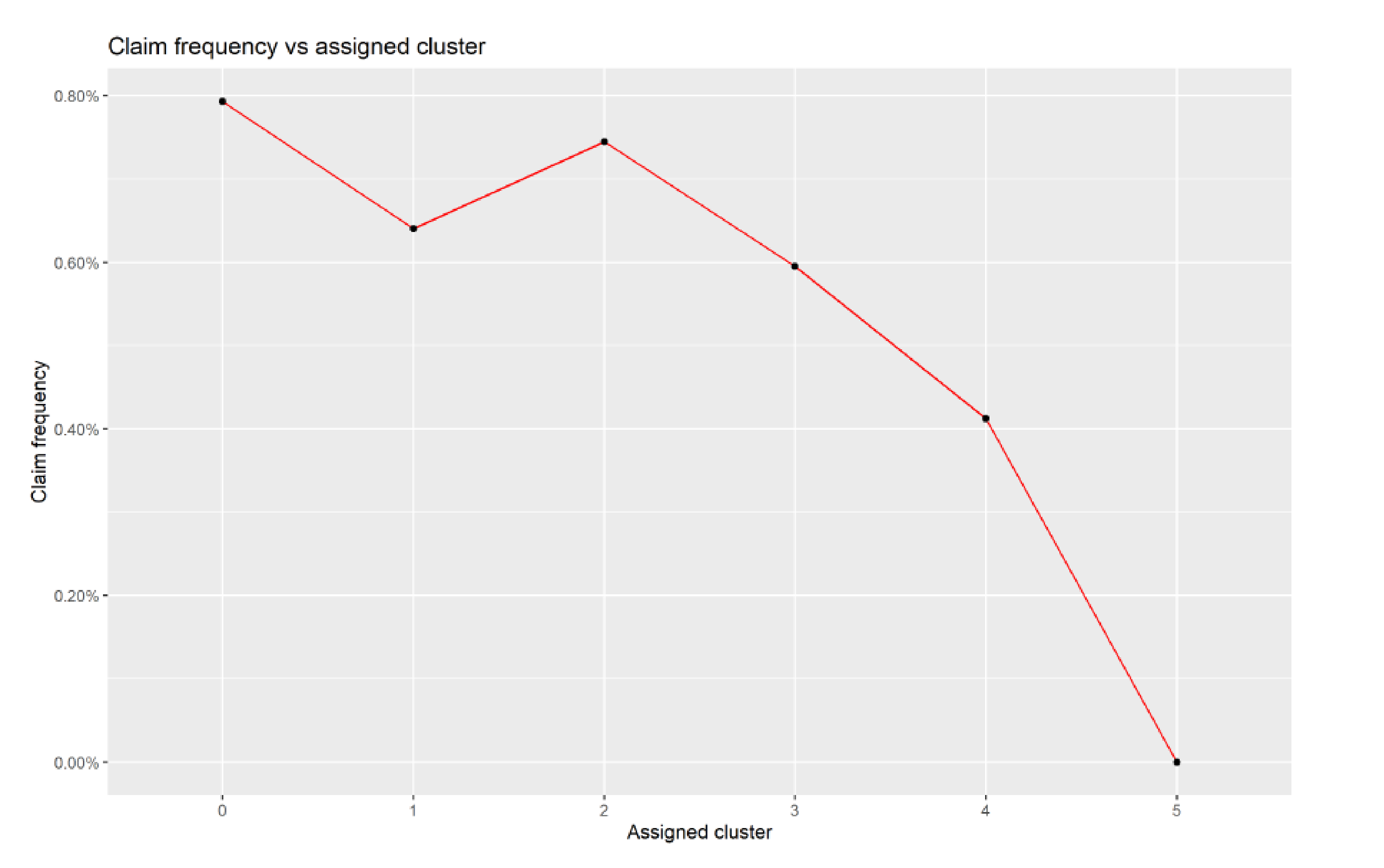
Application of GLRM



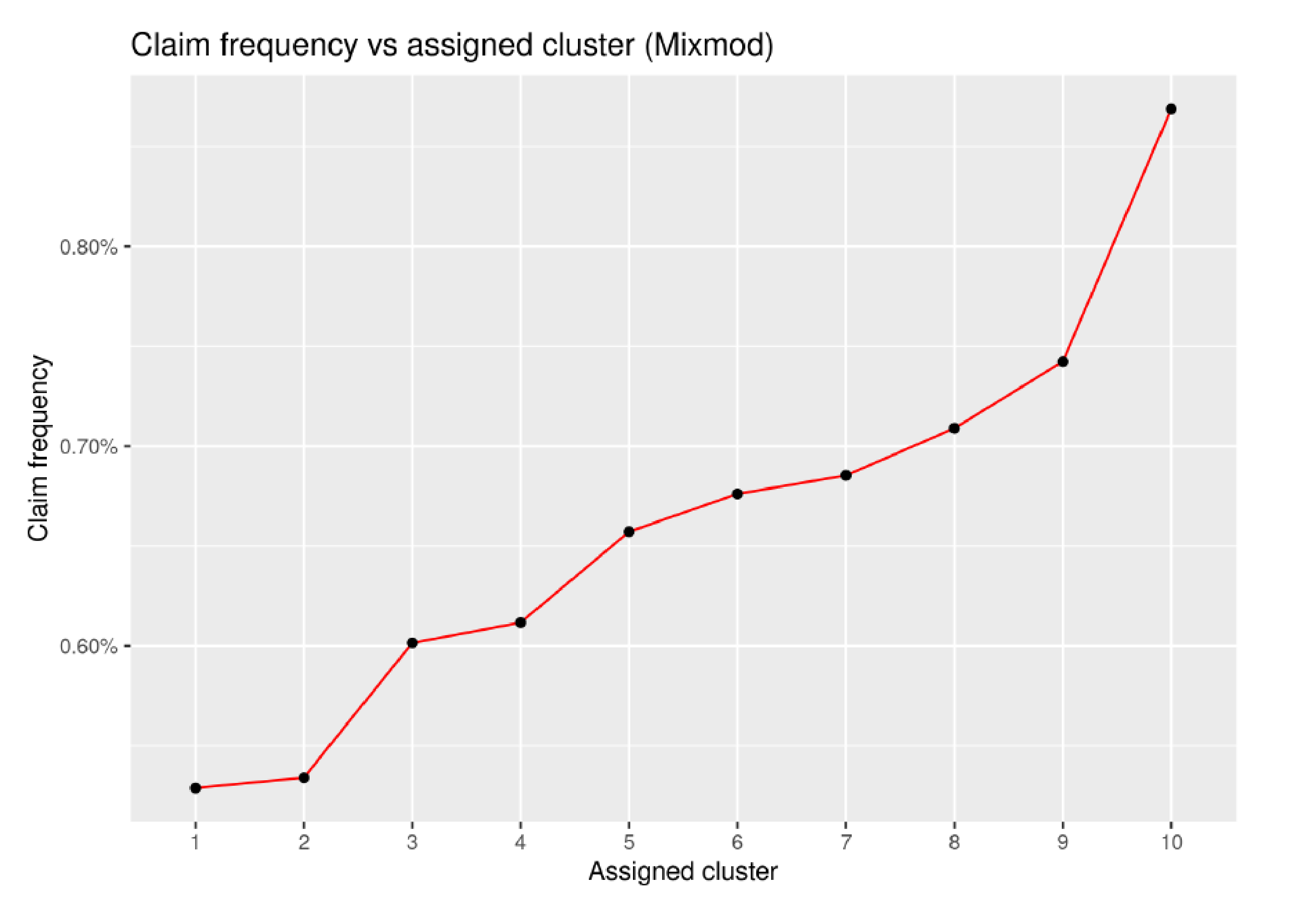
Application of Autoencoders



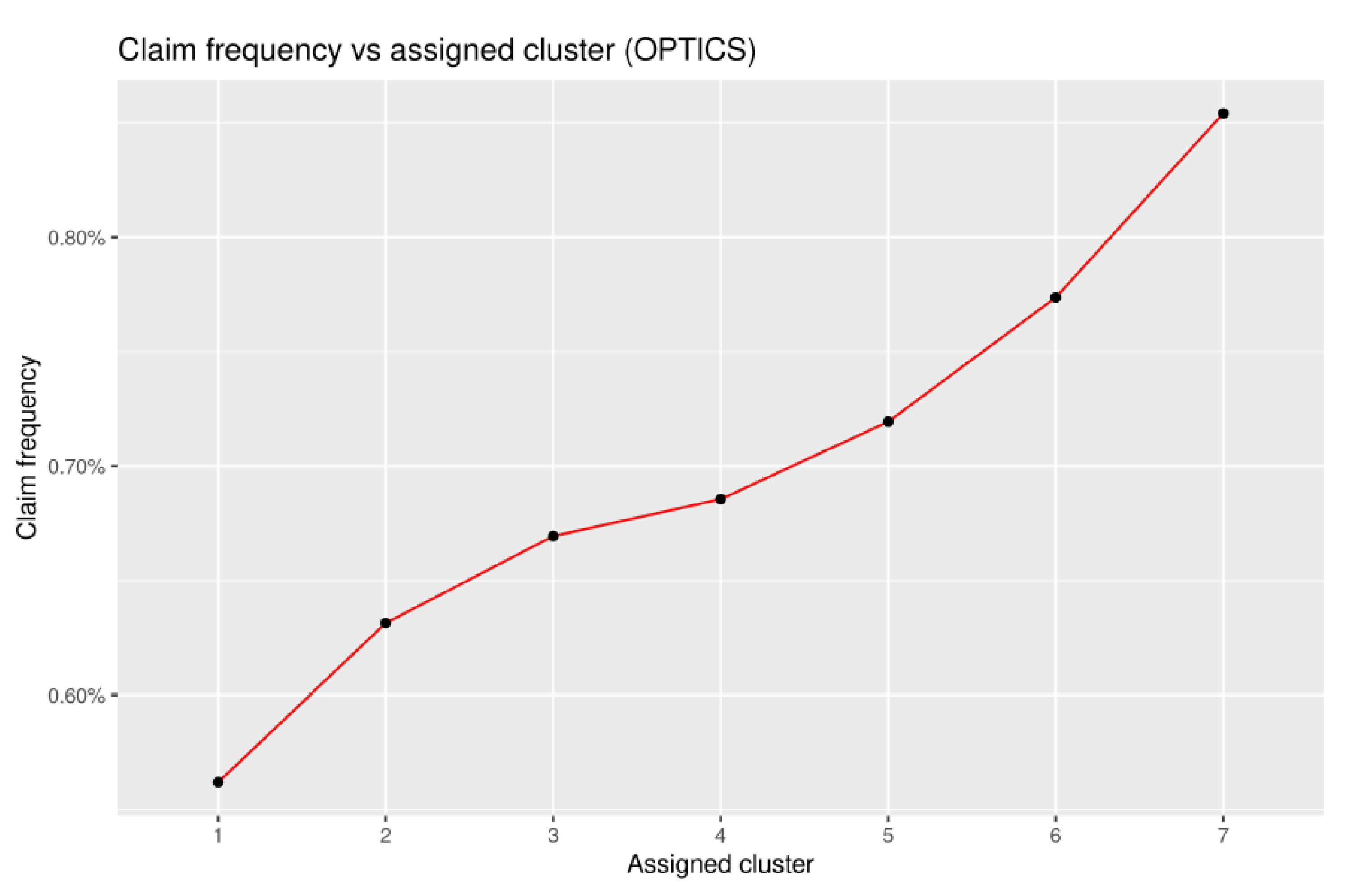
Application of K-Means



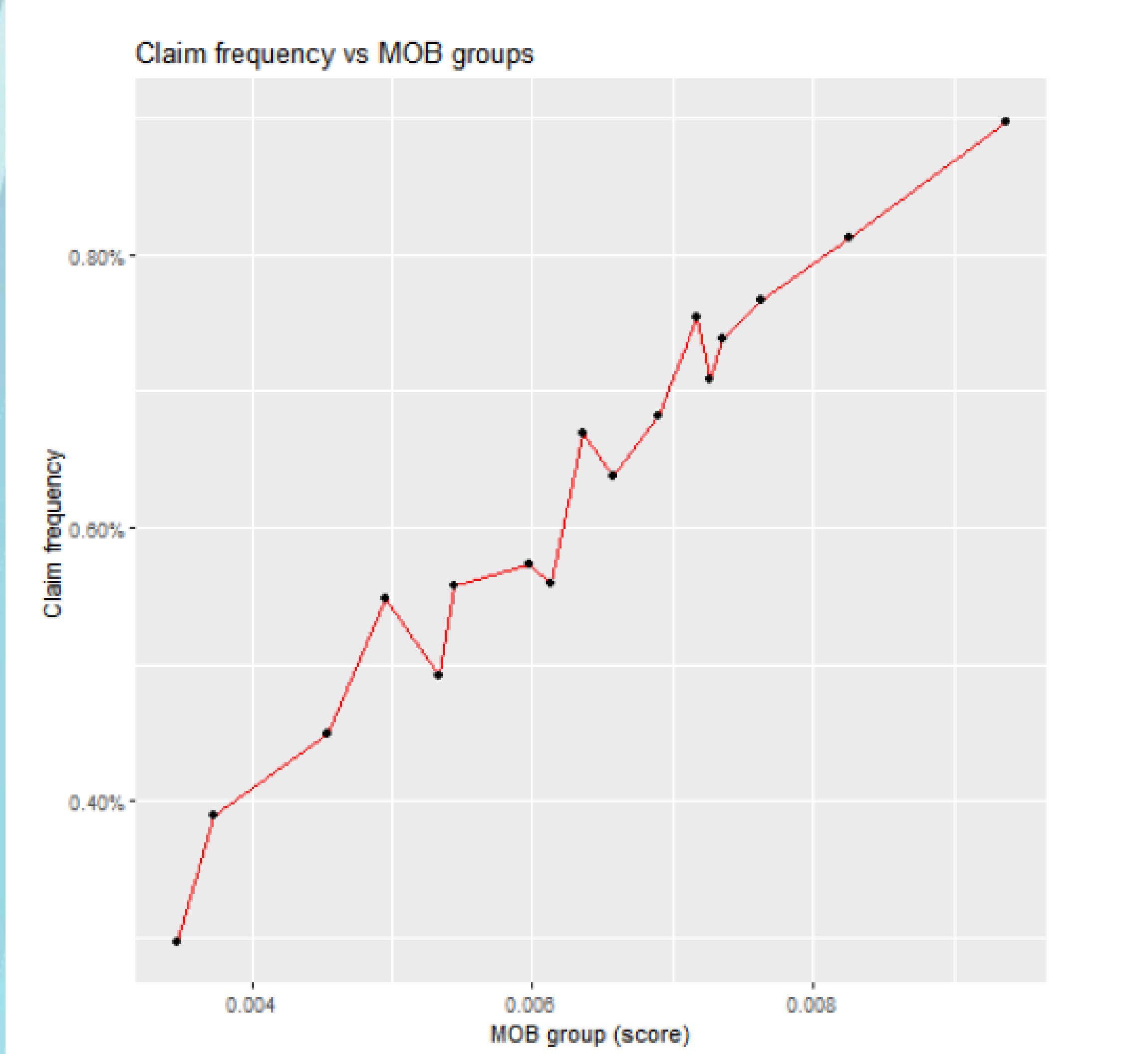
Application of Mixmod



Application of OPTICS



Application of CART



Conclusions

Conclusions

- The **ability to predict the claim frequency** is the main criterion that has been used to rank the different predictive algorithms presented.

- The **Normalized Gini index** has been used to quantify, on the test set, the methods' ability to discriminate vehicle propensity to file claims.

Conclusions

Model	Gini Index
CART	0.468
GLRM	0.334
Autoencoders	0.327
Mixmod	0.314

- The **CART supervised approach** clearly outperforms the other unsupervised methods.
- However, among **unsupervised algorithms,** some latent features of **GLMR** and **Autoencoders** show substantial Gini scores.

Conclusions

- This paper has compared several ML algorithms aiming to define groups of vehicle characterized by **similar loss propensity**, the Vehicle Symbols.
- The predictive power of newer techniques appears to significantly **outperform** older ones.
- Many of these algorithms are very new and **little known** by the predictive modeling practitioners in the insurance industry.
- This research aims to offer an **initial introduction** to the capabilities of such new techniques, in order to encourage more in-depth study by actuaries.
- We believe that it is very beneficial to **explore** these capabilities in the context of actuarial science.

Contacts:

Giorgio A. Spedicato: spedygiorgio@gmail.com

Marco De Virgilis: devirgilis.marco@gmail.com

Casualty Actuarial Society
4350 North Fairfax Drive, Suite 250
Arlington, Virginia 22203

www.casact.org