**Introduction**

Our challenge is to create a machine learning model based on life-insurance customer attributes which will predict that patient’s risk of mortality. Each sample in the training set contains an attribute vector with over 100 features, and a response field that an ordinal measure of risk with 8 levels.

When an individual applies for life insurance, their application includes similar attributes so that the insurer can determine the prospective customer’s premiums and the payout (should the customer die). This process involves assessing the applicant’s risk of mortality. Here we have a predictive analysis problem that can be addressed by countless statistical models - but which one is the best? Solving this problem with machine learning will streamline the application process, give potentially instant feedback and price estimates to the customer, remove overhead for the insurer (by not estimating the risk by hand), and if the model is accurate enough it can save money by predicting risk better than humans can, optimizing profits. Therefore there is a very lucrative business use case for machine learning in life-insurance risk modelling.

This is a complex problem, however. There are many attributes available for consideration, and even then, there are many more factors that decide life expectancy that are not measurable, such as random accidents. The combination of high dimensionality and uncertain classification makes this problem particularly prone to overfitting.

For the midterm submission, we’ve elected to implement a gradient boosted tree. We’ve implemented it using an SKLearn package and achieved 52.207% accuracy.

**Related work**

Galindo, J., & Tamayo, P. (2000). Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications. *Computational Economics*, *15*, 107-143. Retrieved from: <https://link.springer.com/content/pdf/10.1023%2FA%3A1008699112516.pdf>

The above paper assesses risk levels of credit in institutional portfolios. Risk analysis in finance is a critical component of successful investing. Devising risk is not a one-size-fits-all solution, and each speculative asset (e.g. a life insurance plan) has different metrics by which one can determine their risk.

The paper above uses mortgage loans as an example of a single speculative asset for which risk analysis is both important and complex to solve. To understand the concept of credit risk for this asset, imagine that one institution was to purchase a homeowner’s mortgage loan from another. The valuation of that loan is based on the risk of it not being paid back by the lendee. This risk is similar to that of our problem. Although the risk index is a continuous probability instead of an ordinal value (key difference), it *is* determined by 24 attributes such as the size of the loan, whether it is for construction or an existing house, payment history of the loan, overdue balance, etc. These attributes, like the ones in our assignment, consist of both discrete and continuous values (Galindo 119).

In this study, over 9,000 models were built to evaluate their effectiveness at estimating risk of default on mortgage loans. The most effective algorithms, and also the ones mentioned in the paper, include: CART decision-tree models, neural networks, K-nearest neighbor, and probit. The most effective algorithm, an implementation of decision-trees, achieved an 8.31% error rate with only 2,000 training samples (Galindo 107).

*Strengths:*

The breadth of algorithms explored in this study is a key feature of its importance. With 9,000 models built, it is believable that the researchers explored all possible approaches to the problem which existed at the time.

The implementation details of each algorithm (preprocessing strategies, algorithm configurations) are helpful in guiding our approach to this project’s solution. Additionally, the minimal error rate of 8.31% provides an excellent target for our best algorithm implementation.

Finally, the paper explicitly discusses the possibility of applying similar machine learning models to other types of financial risk, implying that risk analysis for life insurance is a viable candidate for machine learning (Galindo 136).

*Weaknesses:*

It is important to note, however, that this study was conducted 18 years ago and since then there have been numerous innovations to the algorithms implemented. Additionally, it gives us no idea how algorithms devised since the year 2000 would perform relative to those studied. The number of potential risk models today is much greater than it was at the time of this study. For example, the algorithm we chose for our midterm submission is a gradient-boosted random forest, whereas the researchers specifically did not attempt boosting in their models (Galindo 135).

Another point of weakness is the size of the dataset utilized in this study. There were only 4,000 samples to work with, however the implementation of the CART decision-tree model (the most accurate) requires an estimated 22,000 samples for optimal performance on this dataset (Galindo 133).

Finally, the dataset in this study has about 1/5th as many attribute fields as the one in our project - so they did not see a need to employ dimensionality reduction, which may be necessary to our optimal solution. High-dimensionality datasets like ours may require more samples for the predictive algorithm to be as accurate as it would be on a low-dimensionality dataset, so we may face issues that these researchers did not encounter/address.

**Implementation**

See attached Code for midpoint implementation

**Result interpretation**

Our results were useful. The highest occuring class in the training set was 8, which occurred between 30% and 40% of the time. So by always guessing 8, you would expect an accuracy under 40%, while out implementation achieved well over 50% accuracy. Our model may also be suffering from overfitting, so we will adjust the maximum depth of the random forest and examine dimensionality reduction in conjunction with other algorithms going forward.