Using Machine Learning to Better Model Long-term Care Insurance Claims

Jared Cummings[a,*], Brian Hartman[a]

[a]*Department of Statistics, Brigham Young University, Provo, UT, USA*

Abstract

Long-term care insurance (LTCI) should be an essential part of a family financial plan. It could protect assets from the expensive and relatively common risk of needing disability assistance. In spite of this, LTCI purchase rates are lower than expected. While there are multiple reasons for this trend, it is partially due to the difficultly insurers have operating profitably as LTCI providers. If LTCI providers were better able to forecast claim rates then they would have less difficulty maintaining profitability. In this paper, we develop several models to improve upon those used by insurers to forecast claim rates. We find that standard logistic regression is vastly outperformed by tree-based and neural network models. More modest improvements can be found by using a neighbor-based model. Of all our tested models, the random forest models were the consistent top-performers. Additionally, simple sampling techniques influence the performance of each of the models. This is especially true for the deep neural net which improves drastically under oversampling. The effects of the sampling vary depending on the size of the available data. To better understand this relationship, we thoroughly examine three states of with various amounts of available data as case studies.

1. Introduction

About half of the 50-year-olds in the United States will spend time in a nursing home before they die (Hurd et al., 2014). About 10% will incur long-term care expenses in excess of $200,000 (Favreault and Dey, 2016). Long-term care insurance (LTCI) should be able to protect members' assets from this expensive risk that is, relative to other insurable risks, common. Surprisingly, only about 10% of individuals over the age of 62 have private LTCI (Braun et al., 2019). There are three main reasons that long-term care purchase rates are lower than expected: public insurance, adverse selection, and difficult profitability for insurers.

Public Insurance. Medicaid offers assistance for nursing home expenses for those who meet a means test. Because of this, when looking at a sample from the Health and Retirement Survey, LTCI purchase rates are lower for those in the bottom income quintile (2%) than for those in the top quintile (20%). In addition to Medicaid, a simple lack of means to pay premiums can also reduce the purchase rates for those with less income. This does not explain why those with the highest incomes don't more commonly purchase LTCI.

Adverse Selection. People understand more about their long-term care risk than is readily observable by an agent or insurer. Finkelstein and McGarry (2006) found that self-reported nursing home entry probabilities are predictive of nursing home use even after controlling for other observable characteristics.

Difficult Profitability for Insurers. When an insurance product is exposed to adverse selection it is more difficult for an insurance company to profitably write. Additionally, most long-term care insurance is purchased from brokers or agents (Thau et al., 2014) whose commission could be equal to or exceed the first year's premium. The underwriting questionaire is long and detailed, increasing costs for the insurer and reducing the number of people willing to apply for LTCI. Rejections are common. Thau et al. (2014) found that 20% of formal applications are rejected. Additionally, Ameriks et al. (2016) found that some affluent individuals are not interested in the products currently available, but would be interested in another product not currently available.

[*]Corresponding author

Email addresses: jared.cummings4@gmail.com (Jared Cummings), hartman@stat.byu.edu (Brian Hartman)

All of the above difficulties in writing LTCI imply that accurate modeling of long-term care risk is essential for individual insurers, the industry, and to protect retiree assets across the country. The difficulty of modeling long-term care risk can be seen in the current loadings and limitations on the policies. In 2000, a representative policy only covered 34% of expected lifetime costs (Brown and Finkelstein, 2007). In 2010, a representative policy covered 66% of lifetime costs (Brown and Finkelstein, 2011). Those two studies also note that the loads on LTCI policies varied from 0.2-0.5 (increased premium above the expected losses, accounting for expenses, commissions, and profit). Those loads are much higher in LTCI than in other industries; e.g. 0.04-0.15 in group medical insurance (Karaca-Mandic et al., 2011) and 0.15-0.25 in life annuities (Mitchell et al., 1999).

Modeling in LTCI takes two main forms, depending on the data available. The first is modeling the population as a whole and their propensity to need long-term care insurance through reductions in their ability to perform the common tasks of daily living. This provides information about the market as a whole and the potential to better understand the disability process more broadly. The second option is to look directly at long-term care insurance industry data to better predict claims directly. This has the advantage of incorporating the unique characteristics of those who purchase long-term care insurance, similar to the differences in the annuitant population in life insurance compared to the entire population.

There is an active and robust stream of literature examining disability transition models. Olivieri and Pitacco (2001) built a model with a single level of disability. Rickayzen and Walsh (2002) developed a multi-state model for the United Kingdom and discussed the impacts of their model on the future need for long-term care. Leung et al. (2004) applied the same model in Australia. Leung (2004) generalized Thiele's differential equation to perform reserve and premium calculations for an LTCI product in Australia. Pritchard (2006) used a seven-state transition model. Stallard (2011) presented a life table approach. Brown and Warshawsky (2013) showed that disability transition rates vary greatly based on initial health status. Fong et al. (2015) used a GLM to estimate a three-state functional disability model. Shao et al. (2017) built a four-state model and used it to price products. Li et al. (2017) incorporated systematic trend and uncertainty. Recently, Sherris and Wei (2020) developed a model both of functional disability and health status and used that model to price a variety of insurance products. They also found that integrating LTC and life annuities in a single product can help reduce systematic uncertainty.

Modeling directly on LTCI industry data is not as prevalent for a few reasons. First, industry data is more difficult to acquire as companies strive to maintain competitive advantage and protect the privacy of their policyholders. Second, industry information is much harder to generalize across companies due to differences in policy characteristics, underwriting practices, and marketing emphases. Lally and Hartman (2016) analyzed industry data from a single large LTCI insurer. The Society of Actuaries has published 6 different reports since 2015 looking directly at the long-term care industry in the United States. In 2015, they published three reports based on a 2000-2011 long-term care intercompany experience study. In January, they published a report describing the data collection and fit GLMs to predict claim incidence, termination, and utilization (Bodnar et al., 2015b). In April, they developed basic tables for LTCI experience (Bodnar et al., 2015a). In July, they produced the policy terminations aggregate databases (Purushotham, 2015). In the subsequent years, they produced reports on persistence (Ho, 2016), table caveats (Society of Actuaries, 2017), and incidence rates over time (Morton and Donato, 2018).

In this paper we build on the work above to develop better models for predicting claim incidence in LTCI. We use a countrywide dataset including data from 12 of the 22 largest LTCI companies in the US (their particular identities are hidden to maintain privacy). We fit several different models to the state-level data to find which best predicts claim incidence. Ideally these models could help insurers reduce premium (by reducing the load), make more steady profits, and help the LTCI market stabilize.

2. Data

Our dataset is developed from the claim incidence spreadsheet accompanying Bodnar et al. (2015b). It contains 14.8 million policy years and 29 variables (Table 1). The amount of data in each state is radically different from the three biggest states (FL-1.3M policy years, CA-1.1M, and TX-820K) to the three smallest

(AK-11K, WY-25K, and WV-28K). This state-to-state variation in data quantity and the inherent differences between each state's regulatory environment led us to consider each state individually. Accordingly, data was separated by state prior to any preprocessing or modeling.

The data, in its form used by Bodnar et al. (2015b), tracked the number of unique claims an individual had in the data development period. A unique claim was defined as a claim not within six months of another claim. For our analysis, we had all of the exposures and claims with the same set of explanatory variables grouped together. We divided them up into single exposure rows to allow for binary predictions, rather than counts. This more closely follows what an actuary would do in practice, predicting whether or not each individual policyholder will have a claim.

3. Methods

We are interested both in preprocessing through training sampling and comparing various classifcia-tion methods. We begin by discussing the training sampling approaches used and then discuss modeling considerations. We then give special emphasis to the architecture of the neural network.

3.1. Training Sampling

Because the proportion of observations that had a claim was consistently low, ranging from 0.007 in Hawaii to 0.05 in Florida, we investigate different sampling approaches on training data and their influence on model fit. We were specifically interested in how simple oversampling and undersampling would impact classification strength measured by a model's area under the ROC curve (AUC). Oversampling is sampling with replacement from the less-represented class (in our case those on claim) until both classes are equally represented. If there are n observations without a claim, then there will be $2n$ observations in this new dataset, half of which had a claim. This necessarily exposes a model to the some or all observations with a claim more than once. Undersampling is a similar process, but instead of sampling from those with a claim a sample is taken from those without a claim without replacement until again both classes are equally represented. If there are k observations with a claim, then there will be $2k$ observations in this new dataset, half of which are sampled from those without a claim. Undersampling removes a potentially large proportion of the training data and many observations without a claim may never be used for model training or holdout evaluation.

3.2. Modeling pipeline

To begin the modeling pipeline, state-segmented data was separated into train and test sets using a 90/10 split. We chose to use cross-validation for models that required a hyperparameter search in order to guard against validation bias as models were tuned. Because of this no explicit partition of the training data was held out for validating these models. At this stage, training data was both oversampled and undersampled. Then, the test set and all three training sets (the original set, the oversampled set, and the undersampled set) were temporarily saved to ensure that all models were fitted with and evaluated on identical data. This facilitated direct comparison between the models after taking into account the sampling treatment of the training set used.

Five types of models were fit for each of the three sampling methods mentioned above, totaling fifteen models for every state. The five model types included logistic regression, K-nearest neighbors (KNN), a gradient boosted forest, a random forest, and a deep neural network (DNN). All models were written in Python, and all but the neural network were implemented using the appropriate functions from the Scikit-Learn package. The DNN was written in Keras with the Tensorflow package for backend computation. For the two tree models, a gridsearch was used to determine optimal hyperparamter values. We considered the logistic regression to be the baseline for model performance.

After any needed hyperparameter tuning, all models were fit on the entire training partition and then evaluated on the test set. AUC scores were taken to compare model performance.

Variable	Description
Gender	Gender of the policyholder
IssueAgeBucket	Issue age of the policyholder, divided into 5 year buckets between ages 50 and 90
IncurredAgeBucket	Age of the policy holder during the policy year, divided into 5 year buckets between ages 50 and 90
IssueYear	Year the policy was issued, divided into 3 year buckets between 1972 and 2011
PolicyYear	Tenure of the policy, between 1 and 37
MaritalStatus	Marital status of the policyholder
PremClass	Premium class of the policyholder, whether preferred, standard, or substandard
UnderwritingType	Underwriting type, whether full or other (accelerated, guaranteed issue, etc.)
CovTypeBucket	Coverage type, whether comprehensive or other
TQStatus	Tax-qualified status, whether qualified, not qualified, or unknown
InflRider	Inflation rider, whether guaranteed purchase option, inflation rider, none, or unknown
RateIncreaseFlag	Indicates whether this policy has ever had a rate increase
NHOrigDailyBenBucket	Original daily benefit for nursing home claims grouped into <100, 100-199, 200+, and unknown
ALFOrigDailyBenBucket	Original daily benefit for assisted living facility claims grouped into <100, 100-199, 200+, and unknown
HHCOrigDailyBenBucket	Original daily benefit for home health care claims grouped into <100, 100-199, 200+, and unknown
NHBenPeriodBucket	Benefit period of the policyholder bucketed into year groups of <1, 1-2, 3-4, 5+, unlimited, and unknown for nursing home claims
ALFBenPeriodBucket	Benefit period of the policyholder bucketed into year groups of <1, 1-2, 3-4, 5+, unlimited, and unknown for assisted living facility claims
HHCBenPeriodBucket	Benefit period of the policyholder bucketed into year groups of <1, 1-2, 3-4, 5+, unlimited, and unknown for home health care claims
NHEPBucket	Elimination period of the policyholder bucketed into 0, 20, 30, 60, 90/100, >100, and unknown for nursing home claims.
ALFEPBucket	Elimination period of the policyholder bucketed into 0, 20, 30, 60, 90/100, >100, and unknown for assisted living facility claims
HHCEPBucket	Elimination period of the policyholder bucketed into 0, 20, 30, 60, 90/100, >100, and unknown for home health care claims
Region	Location consolidated into four regions (Midwest, Northeast, South, West, and unknown)
StateAbbr	Location of the policy by state
Exposure	Exposure in policy years
ClaimCount	Total claims
CountNH	Nursing home claims
CountALF	Assisted living facility claims
CountHHC	Home health care claims
CountUnk	Unknown claims

Table 1: Policy-level covariates

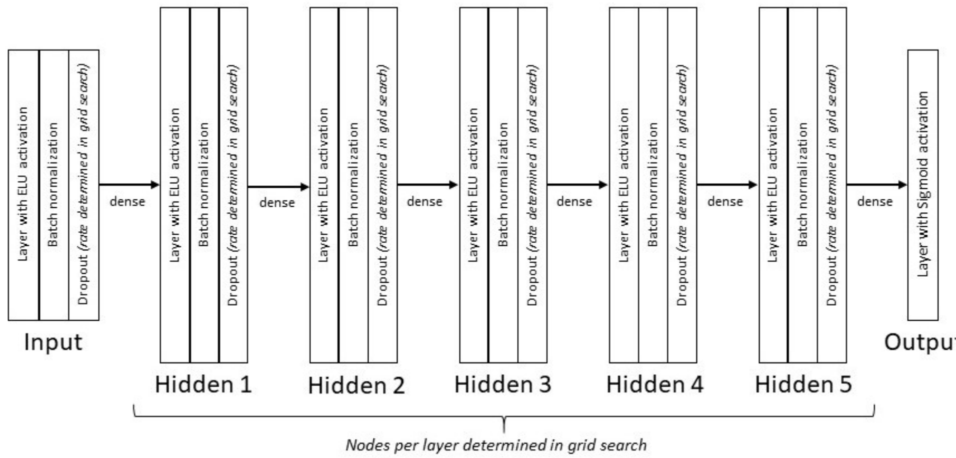

Figure 1: Schematic of Deep Neural Network Model

3.3. Neural Network Architecture

Of special interest is the architecture of the neural network (Figure 1). The size of these kinds of models combined with hyperparameter settings can improve or hurt model performance, so care was taken to construct the model in a way that is consistent with the modeling objectives. With regard to model size, generally measured in the number of parameterized nodes in the network, Delalleau and Bengio (2011) hypothesized that deeper networks are better able to learn hierarchical relationships within the data. Their hypothesis was confirmed more recently by Mhaskar et al. (2017). The implication of this for practitioners is that deep neural networks more efficiently create and use sub-representations of data relationships than shallow, broader networks with the same number of parameterized nodes. In accordance with this, we built a model with 5 hidden layers, to keep the model deep enough for our classification task, but also allow the number of nodes per hidden layer to be tuned through a hyperparameter grid search. This allowed the total model size to vary without losing the learning benefits that deeper networks provide.

To keep the gradient tractably small, batch normalization was applied after every layer except for the output layer. Exponential linear units (ELUs), introduced by Clevert et al. (2015), were used as the activation for those same layers, selected for their near-zero mean and computational advantages. To be consistent with the original ELU implementation, He initialization, first used by He et al. (2015), was used for initial weights. All biases were initialized at zero. A sigmoid function was used as the activation function for the output layer so the output could be interpreted as the probability that an observation would go on claim given the data features.

Gradient descent on the network was performed using the Adam optimizer introduced in Kingma and Ba (2014) for its computational speed, though this did come at a potential performance trade-off as Wilson et al. (2017) did show that this optimizer will never reach the global optimum and risks severe over-fitting. To reduce this risk of over fitting, dropout regularization, introduced in Hinton et al. (2012), was employed after batch normalization. The dropout rate was determined for each model by grid search.

4. Results

After the models performance was assessed, some trends emerged. The first was a confirmation of an accepted fact: with more data, models are more powerful. Every model performed better when more data was available for training. The neural network, random forest, and KNN algorithms were the most sensitive to the increase in data quantity. Gradient boosted forest and logistic regression were somewhat less sensitive, but had the same performance bump.

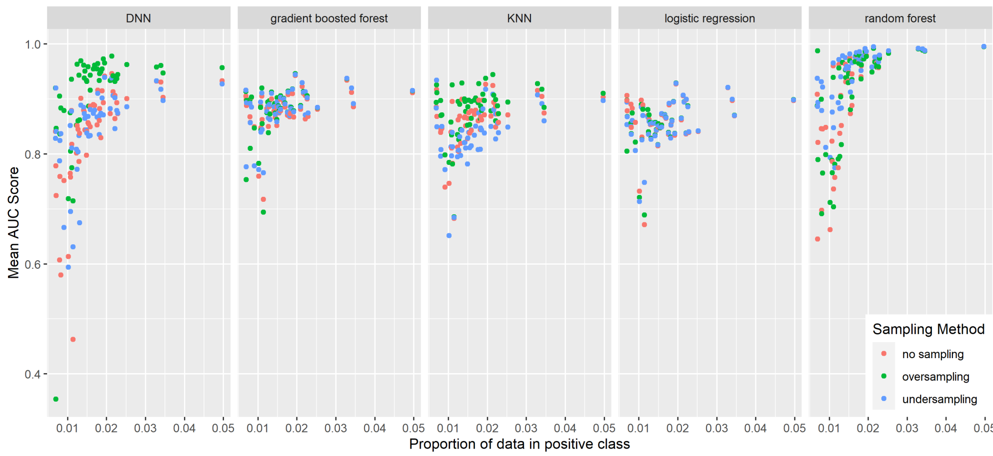

Figure 2: Model AUC Performance Against Positive Class Prevelance, Faceted by Model

Also notable was that different models respond differently to different sampling approaches (Figure 2). Again, the gradient boosted forest and logistic regression were least sensitive to the changes. Logistic regression, however, had fewer low-performing models when oversampling was used, but by a narrow margin. The random forest algorithm preferred undersampling when the positive class was less represented (in number of claims per policy), and as the representation increased undersampling and oversampling performed similarly. Both the KNN and DNN vastly preferred oversampling.

The effectiveness of sampling varied strongly with the claim prevalence. When less than 1.5% of the data had claims all models were less able to able to discriminate between observations regardless of sampling method. This effect was most pronounced in the models that were top-performers when the positive class made up a larger share of the observations. As an example, no neural network trained using oversampled data when more than 1.5% had claims had an AUC score lower than 0.9, while scores when less than 1.5% had claims varied wildly from above 0.9 to below 0.4. The variability of AUC scores for similarly represented data did not increase as sharply when data was undersampled or imbalanced, though for the DNN these models consistently underperformed compared to models trained with oversampled data. No analysis was done to determine what data attributes other than claim proportion accounted for the variation.

With the exception of some states with few claims, the random forest model tended to outperform all other models (Figure 3). The DNN was similarly dominant when oversampling was used, but when the training data was undersampled or imbalanced the network remained a strong performer, but did not distinguish itself from the other three models. The consistent outperformance of the gradient boosted forest by the random forest suggests that this data problem is better approached by increasing the number of times the data is seen by a model and not simply by optimizing a gradient descent. The success of the oversampled DNN over other models also points to a need to find deeper representations of intervariate relationships for which ensembling and deep learning are better suited.

5. Case Studies

Taking a deeper look at model performance on data of specific states helps us to understand the nuances of the model and sampling method interplay. We look closer at models trained on data from Montana, Kentucky, and Virginia, approximately representing the first quartile, median, and third quartile, respectively, for both number of observations and frequency of claims. We begin with Kentucky.

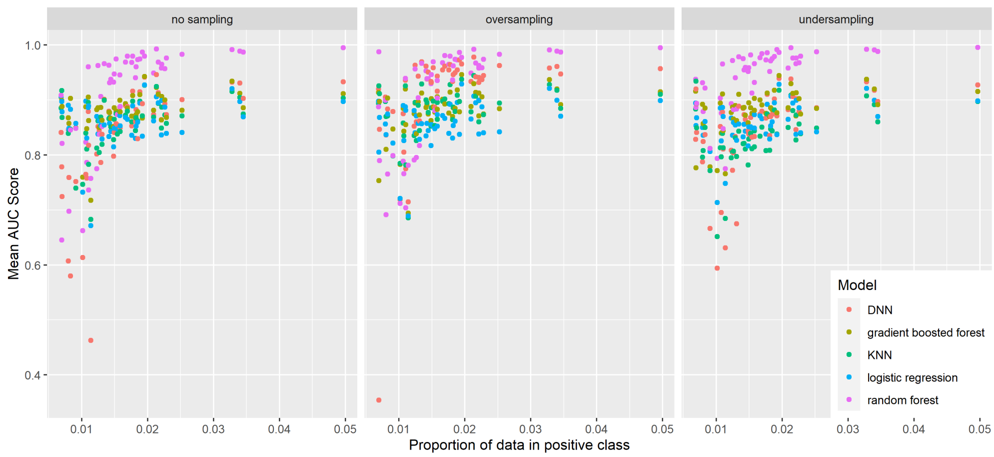

Figure 3: Model AUC Performance Against Positive Class Prevelance, Faceted by Sampling Method

5.1. Kentucky

After all of the models are fit and evaluated as we describe in the methods section, we can take a look at the AUC scores (Table 2). A cursory overview shows that a random forest model (RF) with undersampled training data is the top performer with an AUC score of 0.932. It is followed by a deep neural network (DNN) and random forest with oversampled training data, scoring 0.916 and 0.903 on the AUC metric respectively. The other models have less impressive results, with a KNN model using undersampling being the lowest performer (0.815 AUC).

Sampling	Model	AUC
undersample	RF	0.932
oversample	DNN	0.916
oversample	RF	0.903
imbalanced	RF	0.888
oversample	GBF	0.883
undersample	GBF	0.881
imbalanced	GBF	0.878
oversample	KNN	0.865
imbalanced	Logit	0.853
oversample	Logit	0.852
undersample	Logit	0.849
imbalanced	DNN	0.849
imbalanced	KNN	0.843
undersample	DNN	0.833
undersample	KNN	0.815

Table 2: AUC scores for Kentucky models

To get a different view of model performance, we can consider the ROC curves of these models (Figure 4). From these plots, we can see that there is something going on with the random forest model when no sampling is applied (the data is left imbalanced) and when the training data is oversampled. The totally vertical segment along the y-axis of these graphs suggests that there is a high proportion of positive-class predictions that are being correctly identified with an equal probability at or near 1. The long, straight segments on both plots as the false positive rate approaches 1 indicates that there might also be a sizable segment of our positive-class observations being misidentified entirely and assigned an equal probability near

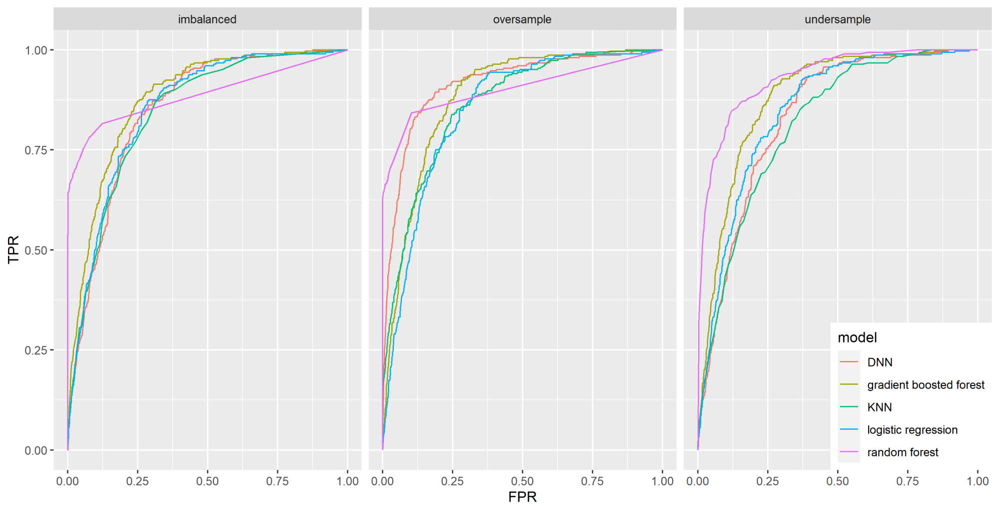

Figure 4: ROC Curve of Kentucky Models, Faceted by Sampling Method

0. A different view shows that the density of predicted probabilities sits on the tails for the imbalanced and oversampled random forests, confirming the behavior seen in the ROC curves (Figure 5)

The bimodal distributions of claim predictions seen in the imbalanced and oversampled random forest models indicate that the models have problems with overfitting. Because we do not see this problem in the undersampled random forest, the culprits for the overfitting are the non-claim observations present in imbalanced and oversampled training sets, but not the undersampled one. The bimodal appearance of the oversampled DNN predictions also suggests the same kind of overfitting. From these modeling results we can hypothesize that there is a small subset of LTC claims that look nearly identical to non-claim observations. The presence of these observations gives the undersampled random forest an advantage over the other models because the algorithm already has a strong ability to discriminate and is not likely to be exposed to the non-claim observations that were contributing to overfitting. Thus, we are able to conclude not just that the undersampled random forest is the best model for Kentucky data, but also we are able to gain an intuition as to why.

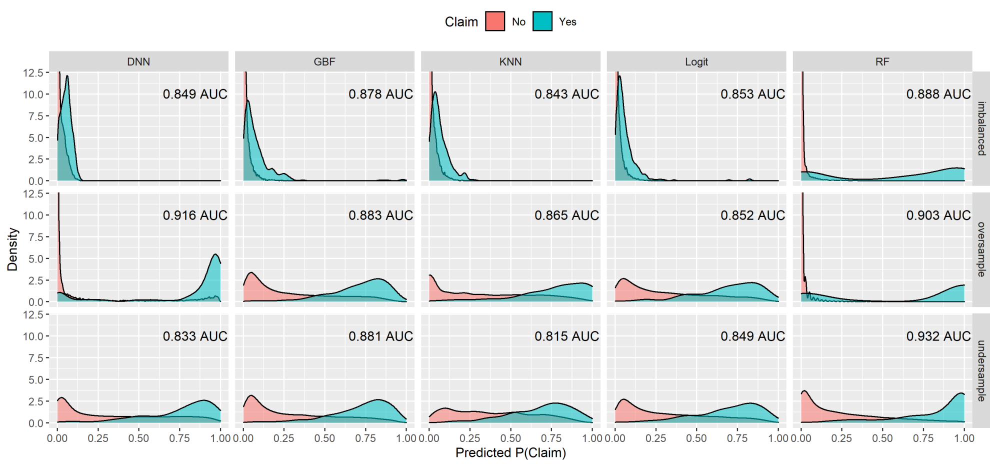

Figure 5: Density of Model Predictions for Kentucky, Faceted by Sampling Method and Model

5.2. Montana

The data for Montana falls at roughly the first quartile of states when ordered by quantity of data. Looking at its AUC curve, we can get a clearer picture of the ways in which a smaller data set influences model performance (Figure 6).

Immediately apparent is the same problem of random forest overfitting for both imbalanced and oversampled training data. When data is left imbalanced, the gradient boosted forest model also begins to overfit, but not to the same extent. We can understand this as being caused by the lower number of claims present in this data set. Fewer observations means that the subtleties that distinguish a claim from a non-claim are not learned.

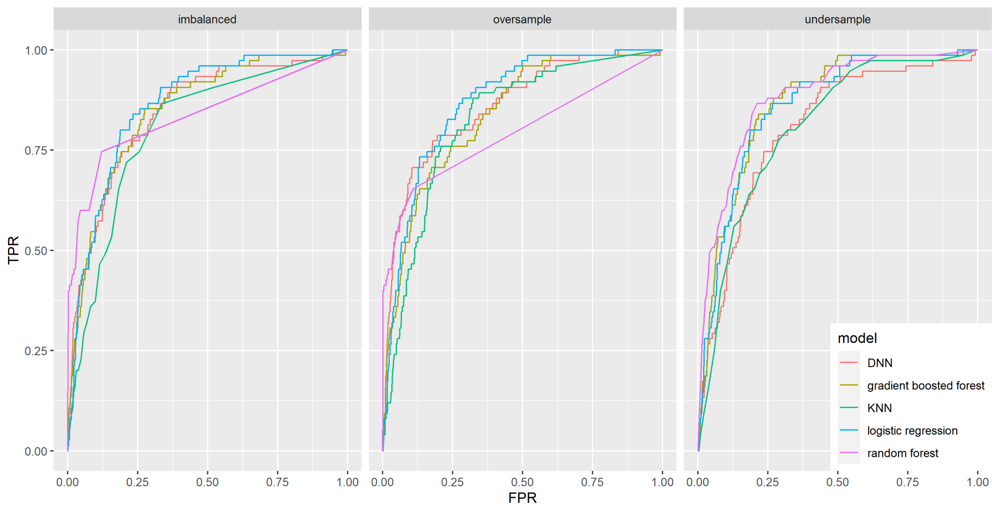

Figure 6: ROC Curve of Montana Models, Faceted by Sampling Method

Looking at Montana's predictive densities, it is apparent that the oversampled DNN overfits more than the same modeling setup did when trained on Kentucky data (Figure 7). The weakness of the network when trained on undersampled data is also exacerbated by having less data to learn from. Overall, the same behaviors are present in Montana as were in Kentucky, only they are more dramatic.

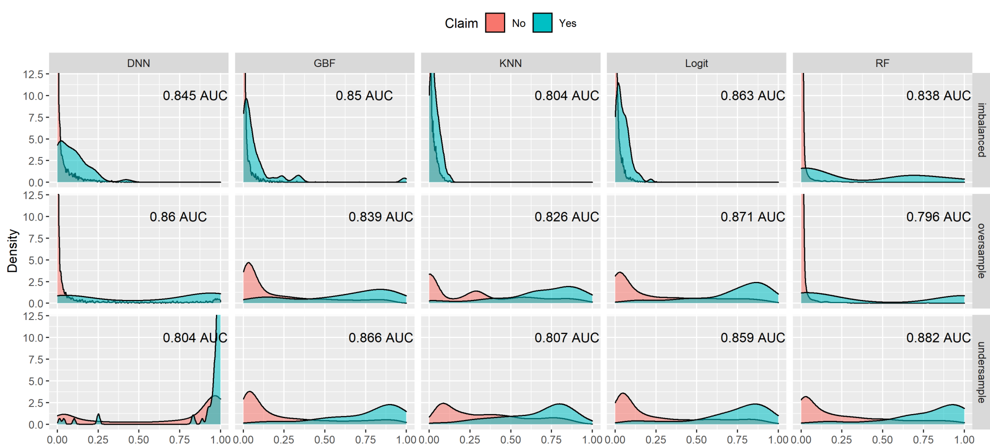

Figure 7: Density of Model Predictions for Montana, Faceted by Sampling Method and Model

5.3. Virginia

Virginia is representative of the top quartile of states by data quantity. So, by comparing Virginia to Kentucky, we get a different story: an understanding of how more data influences the observed modeling behaviors.

With a first glance at the AUC, it seems as though the problems with the random forests have been alleviated (Figure 8). In fact, the random forests all seem to do incredibly well, regardless of data sampling. A look at the modeling densities confirms that the models are learning to discriminate well and are using the full output space to make predictions (Figure 9). A close inspection does show that there is still a small mode for the imbalanced and oversampled models where observations on claim are predicted with probabilities near 0. It is not the case that the overfitting problem has been entirely removed by increasing the data quantity, but it has been significantly reduced so as to be nearly negligible. The same is true for the overfitting of the oversampled DNN, which is still present in the slightest way, but doesn't retain practical importance. Overall, overfitting is alleviated across the board by having more training data available. However, it is still present enough to give the undersampled random forest the edge over the other models under consideration.

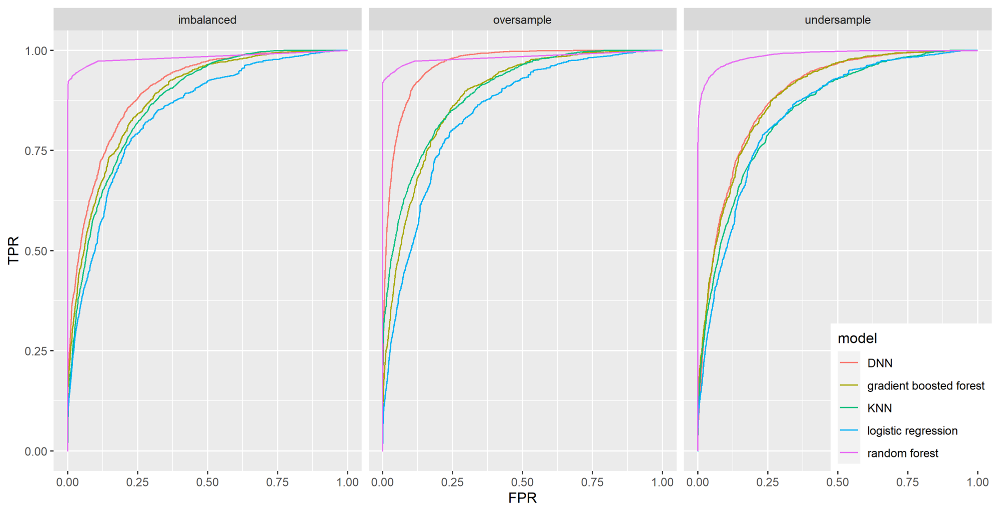

Figure 8: ROC Curve of Virginia Models, Faceted by Sampling Method

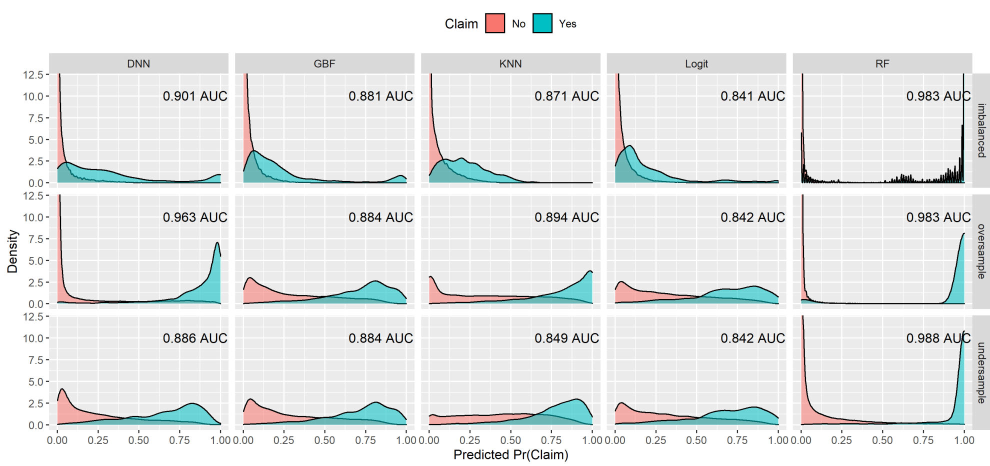

Figure 9: Density of Model Predictions for Virginia, Faceted by Sampling Method and Model

6. Conclusion

Long-term care insurance is an important and underutilized aspect of household asset protection. Improved profitability would encourage more companies to enter the market, improving the product offerings and prices for the consumers. This paper seeks to improve that profitability by improving the claim models in LTCI. There is a large body of work modeling the disability process, but much less at the actual LTCI policy-level.

We compared five different models (logistic regression, K-nearest neighbors, gradient-boosted forests, random forests, and deep neural networks) and three different methods of training sampling (oversampling, undersampling, and using the data as is). We found that the random forest outperformed all the other methods. The deep neural networks were the next best, though this was subject to the quantity of available training data. As both K-nearest neighbors and deep neural networks display dramatic performance improvements with more data, oversampling was preferred especially when the states were smaller. For the other three models, sampling method did not have a consistent impact whether positive or negative.

We also provided an in-depth look at three states, Montana, Kentucky, and Virginia. We showed that in states with a small amount of available data (Montana) and a medium amount (Kentucky) a random forest seems to overfit in the imbalanced or oversampled settings. That problem nearly disappears in when a large amount of data is available for training (Virginia).

Future work in LTCI-specific modeling could include further exploration of the other random aspects of the policies, like lapse and termination rates. There are also many unsolved questions in product design with the life and LTC combination products showing strong promise. There is further potential in improving the models in this paper. For example this could be done by explicitly accounting for spatial and/or temporal correlations or incorporating a hierarchical structure in the data.

7. Bibliography

Ameriks, J., Briggs, J., Caplin, A., Shapiro, M.D., Tonetti, C., 2016. The long-term-care insurance puzzle: Modeling and measurement. Technical Report. National Bureau of Economic Research.

Bodnar, V., Williams, B., Morton, M., Wood, T., 2015a. Long Term Care Experience Basic Table Development. Technical Report. Society of Actuaries.

Bodnar, V., Wood, T., Morton, M., 2015b. Long Term Care Intercompany Experience Study – Aggregate Database 2000-2011 Report. Technical Report. Society of Actuaries.

Braun, R.A., Kopecky, K.A., Koreshkova, T., 2019. Old, frail, and uninsured: Accounting for features of the us long-term care insurance market. Econometrica 87, 981–1019.

Brown, J., Warshawsky, M., 2013. The life care annuity: a new empirical examination of an insurance innovation that addresses problems in the markets for life annuities and long-term care insurance. Journal of Risk and Insurance 80, 677–704.

Brown, J.R., Finkelstein, A., 2007. Why is the market for long-term care insurance so small? Journal of public Economics 91, 1967–1991.

Brown, J.R., Finkelstein, A., 2011. Insuring long-term care in the united states. Journal of Economic Perspectives 25, 119–42.

Clevert, D.A., Unterthiner, T., Hochreiter, S., 2015. Fast and accurate deep network learning by exponential linear units (elus). Under Review of ICLR2016 (1997) .

Delalleau, O., Bengio, Y., 2011. Shallow vs. deep sum-product networks, in: Shawe-Taylor, J., Zemel, R.S., Bartlett, P.L., Pereira, F., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 24. Curran Associates, Inc., pp. 666–674. URL: http://papers.nips.cc/paper/4350-shallow-vs-deep-sum-product-networks.pdf.

Favreault, M., Dey, J., 2016. Long-term services and supports for older americans: Risks and financing research brief. washington, dc: Assistant secretary for planning and evaluation; 2015.

Finkelstein, A., McGarry, K., 2006. Multiple dimensions of private information: evidence from the long-term care insurance market. American Economic Review 96, 938–958.

Fong, J.H., Shao, A.W., Sherris, M., 2015. Multistate actuarial models of functional disability. North American Actuarial Journal 19, 41–59.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. IEEE International Conference on Computer Vision (ICCV 2015) 1502. doi:10.1109/ICCV.2015.123.

Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2012. Improving neural networks by preventing co-adaptation of feature detectors. ArXiv abs/1207.0580.

Ho, C., 2016. U.S. Long-Term Care Insurance Persistency. Technical Report. LIMRA and The Society of Actuaries.

Hurd, M., Michaud, P., Rohwedder, S., Wise, D., 2014. Discoveries in the economics of aging. University of Chicago Press 81, 114.

Karaca-Mandic, P., Abraham, J.M., Phelps, C.E., 2011. How do health insurance loading fees vary by group size?: implications for healthcare reform. International Journal of Health Care Finance and Economics 11, 181.

Kingma, D., Ba, J., 2014. Adam: A method for stochastic optimization. International Conference on Learning Representations .

Lally, N.R., Hartman, B.M., 2016. Predictive modeling in long-term care insurance. North American Actuarial Journal 20, 160–183.

Leung, E., 2004. A multiple state model for pricing and reserving private long term care insurance contracts in Australia.

Leung, E., et al., 2004. Projecting the needs and costs of long term care in australia. Australian Actuarial Journal 10, 343.

Li, Z., Shao, A.W., Sherris, M., 2017. The impact of systematic trend and uncertainty on mortality and disability in a multistate latent factor model for transition rates. North American Actuarial Journal 21, 594–610.

Mhaskar, H., Liao, Q., Poggio, T., 2017. When and why are deep networks better than shallow ones?, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI Press. p. 2343–2349.

Mitchell, O.S., Poterba, J.M., Warshawsky, M.J., Brown, J.R., 1999. New evidence on the money's worth of individual annuities. American economic review 89, 1299–1318.

Morton, M., Donato, R., 2018. Measuring Long-Term Care Insurance Incidence Rates Over Time. Technical Report. Society of Actuaries.

Olivieri, A., Pitacco, E., 2001. Facing ltc risks, in: Proceedings of the 32nd astin colloquium, Washington.

Pritchard, D., 2006. Modeling disability in long-term care insurance. North American Actuarial Journal 10, 48–75.

Purushotham, M., 2015. Long Term Care Intercompany Experience Study – Policy Terminations Aggregate Databases 2000-2011 Report. Technical Report. Society of Actuaries.

Rickayzen, B.D., Walsh, D.E., 2002. A multi-state model of disability for the united kingdom: implications for future need for long-term care for the elderly. British Actuarial Journal 8, 341–393.

Shao, A.W., Sherris, M., Fong, J.H., 2017. Product pricing and solvency capital requirements for long-term care insurance. Scandinavian Actuarial Journal 2017, 175–208.

Sherris, M., Wei, P., 2020. A multi-state model of functional disability and health status in the presence of systematic trend and uncertainty. North American Actuarial Journal 0, 1–23. URL: https://doi.org/10.1080/10920277.2019.1708755, doi:10.1080/10920277.2019.1708755, arXiv:https://doi.org/10.1080/10920277.2019.1708755.

Society of Actuaries, 2017. Caveats for Use of Long Term Care Experience Basic Tables. Technical Report.

Stallard, E., 2011. Estimates of the incidence, prevalence, duration, intensity, and cost of chronic disability among the us elderly. North American Actuarial Journal 15, 32–58.

Thau, C., Helwig, D., Schmitz, A., 2014. 2014 long term care insurance survey. Broker World 13.

Wilson, A., Roelofs, R., Stern, M., Srebro, N., Recht, B., 2017. The marginal value of adaptive gradient methods in machine learning .