A new military Retention Prediction Model: machine learning for high-fidelity prediction

James Bishop* Michael Guggisberg* Julie Lockwood Pechacek*
Alan Gelder* Cullen Roberts* Joe King*

Abstract
The Department of Defense (DoD) is the largest employer in the United States of America with 2.15 million service members in 2019. The DoD must anticipate retention of service members to effectively perform its mission. A team from the Institute of Defense Analyses began development of the Retention Prediction Model (RPM), an application of a feed-forward neural network on a novel dataset. The RPM is a tool that predicts retention of service members in the DoD and can be used to inform DoD leadership about anticipated retention at the service member level. Prediction on an out-of-sample set results in a concordance no worse than 0.78 for any given year or 0.73 for the restricted mean survival time.

Key Words: machine learning, neural network, survival, retention, military

1. Introduction

The Department of Defense (DoD) is the largest employer in the United States of America with 2.15 million service members in 2019.[1] The DoD must have a sufficient number of service members in the correct positions and anticipate retention to effectively perform its mission. Thus, the DoD closely monitors the recruiting and retention of its service members.[12, 19, 26] However, it can be difficult to make strategic personnel decisions without high resolution, high accuracy retention forecasts. High quality retention forecasts could provide DoD leadership with the information to decide priorities in recruiting and staffing allocation.

Machine learning tools (including neural networks) have been deployed in many organizations and industries to predict retention and turnover of employees.[6, 20, 21, 23, 24, 27] More generally, machine learning has been broadly used throughout human resources management to determine staffing, development, and performance management of employees.[25]

Our team at the Institute for Defense Analyses (IDA) was tasked with building a tool to provide retention forecasts for DoD service members. We trained a feed-forward neural network on rich administrative records from 2000 to 2018 that can produce conditional survival probabilities up to 18 years in the future for every active-duty service member in the DoD. The model currently in development is called the Retention Prediction Model (RPM). The RPM yields a concordance index no less than 0.78 for any given time horizon and 0.73 for the restricted mean survival time (RMST) on an out-of-sample subset of the data. The feed-forward neural network was chosen after considering alternative models such as XGBoost, proportional hazards, random forest, and logistic regression.[4, 8]

For over three decades, the DoD has used a structural economic model called the Dynamic Retention Model (DRM) to produce "what-if" counterfactual predictions at the subpopulation level to inform policy decisions.[2, 7, 11, 17] The RPM serves a very different purpose - providing in-sample predictions at the service member level under the current

*Institute for Defense Analyses, 4850 Mark Center Drive, Alexandria, VA, 22311

policy regime. In further work, we intend to complement the RPM with causal inference methods and thereby inform policy changes.

Employing a neural network grants the RPM extraordinary flexibility over alternative survival modeling methods. Whereas a proportional hazards model would assume the effect of features is constant over any time interval and a parametric hazards model would assume the survival probabilities lie on a curve from a pre-specified parametric family, the neural network allows arbitrary nonlinear and interactive relationships among features and outputs. For example, a simpler model would not be able to capture an effect of age on career duration being larger in the Navy than the other services without adding the proper interaction term. While the researcher could add an interaction term between age and Navy membership, the researcher would need to do so consciously, in advance, and for all sufficiently plausible interactions. The RPM has the flexibility to capture arbitrarily complex interactions without researcher prescience.

The primary drawback to such flexibility is the risk of overfitting (capturing relationships that do not generalize to new data). We mitigate overfitting through dropout layers, embedding regularization, and employing a shallow network specification. To estimate the extent of overfitting, we compare model performance on the training and test sets over training epochs.

1.1 How and Why People Leave

The goal of the RPM is to predict retention regardless of reason. However, understanding reasons why service members leave the DoD is important for building a model with strong predictive performance. A priori engineering features that are likely to predict retention can improve neural network performance.

Individuals can join the DoD through enlisted or officer routes. Enlisted service members sign a fixed-year contract that obligates them to a specified number of years of active duty service. A common contract length is four years. As the contract nears its end, the service member and the service jointly decide whether to renew the contract. If the service member or the service decide not to renew the contract, then the service member is released from active duty.

Individuals who join as officers, either commissioned or warrant, typically serve a fixed year initial contract and then continue to serve until the service member or the service decide to end the relationship. The initial contract length for commissioned officers is commonly three or four years.

There are numerous other ways that service members can leave the DoD at contract end points or intra-contract. Service members can be separated due to medical reasons, such as physical injury or mental illness. Service members can also be separated due to disciplinary or legal issues.

2. Data

We use annual observations from a 5% random sample of all active duty members of the DoD from March 2000 through March 2018.[1] The outcome of interest for a given person-year observation is the number of consecutive future years observed for the same service member. We observe 67,423 service members in 2018 but do not use those person-years to train the model, since the outcome value for all 2018 observations is censored at zero and therefore uninformative. From March 2000 through March 2017, our sample contains

[1] We use a 5% sample to reduce training time and increase the number of model iterations.

1,281,471 observations of 216,037 unique service members. We reserve 75% of service members for model training and the remainder for model testing.

We use 368 features on demographics, career, pay, and family. Features can be either numeric (real-valued) or categorical. We identify a feature as numeric if it is a date (converted to Julian) or has more than 1,024 unique numeric values; 53 features, including date of birth, date of initial entry to the uniformed services, and various types of pay, meet this criterion.[2] The neural network can learn contract renewal periods because contract end dates are provided as a feature. We identify all other features as categorical. Categorical features include Service (Army, Navy, Air Force, or Marine Corps), assigned unit, and occupation. The criteria classify low-cardinality but naturally numeric features as categorical, such as number of dependents and Armed Forces Qualification Test percentile. This choice offers additional flexibility for the model to capture non-linear relationships.

Categorical features are encoded by mapping each unique value (including missing values) to a unique whole number and are then passed through an embedding layer (discussed in the next section). Numeric features are min-max normalized to [-0.5, 0.5]. Missing numeric values are set to -1 after normalization.[5, p. 102]

The entry and exit of service members from active duty service within the time frame of the data produces an unbalanced panel. Of all unique service members, 37.9% are observed in the most recent year of data and are thus right-censored. Right-censorship is a critical complication of our research objective and requires survival modeling to address.

Forty percent of service members are observed in the first year of data, many of whom were on active duty in earlier, unobserved years. This left-truncation impairs our ability to engineer new features that depend on career history, such as number of deployments, but does not impair our computation of the outcome.

3. Methods

We train a neural network that maps a given service member to a service member survival curve represented by an 18-element vector of annual conditional survival probabilities. Element t of the output vector represents the probability that the given service member remains on active duty for at least t additional consecutive years. To address right-censorship of the outcome, we use the survival loss function of Gensheimer and Narasimhan when training the neural network.[9]

The neural network consists of a set of parallel embedding layers, one for each categorical feature, followed by two consecutive sets of maxout and 25% dropout layers, then a densely connected maxout layer, followed by a densely connected sigmoid output layer.[10] Each embedding layer outputs a one-dimensional array. Thus each embedding layer is a map from the set of natural numbers to the reals. We use one-dimensional embeddings instead of higher dimensional embeddings for computational efficiency. We do not find embeddings of greater dimensionality improve performance. We impose batch normalization immediately prior to each maxout layer and the output layer. Each maxout layer has 256 maxout units and the activation of each unit is the maximum of four linear activations.

We use the AMSGrad variant of the Adam optimizer to train the neural network with a learning rate of 0.001.[16, 22] We impose regularization on each embedding layer equal to twice the sum of the squared elements of the embedding vector. We train on batches of 512 observations, randomly sampled without replacement, until test set loss does not improve for four consecutive epochs. We then restore the model with the lowest test set loss (from the fourth previous epoch), "freeze" the embedding layers (i.e., stop the embedding vectors

[2]Missing date of initial entry values are imputed with the date of the first observed record for service members first observed later than January 2000.

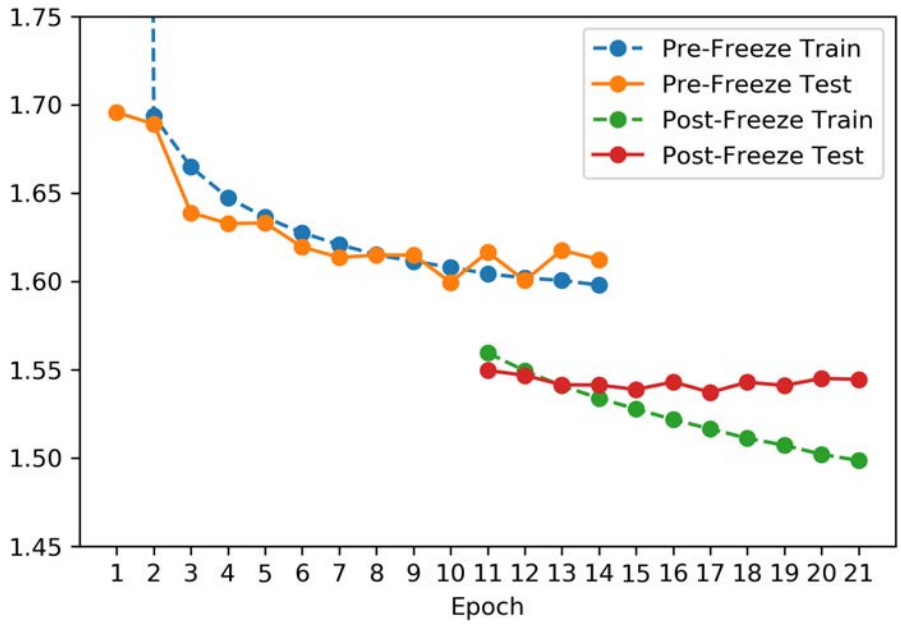

Figure 1: Training and test set loss by epoch

from updating in the upcoming training session), and continue training until test set loss again does not decrease for four consecutive epochs. We then restore the model with the lowest test set loss.

4. Results

Figure 1 presents the loss over epochs as the model is trained. The model trained for ten epochs before failing to decrease test set loss for the next four epochs. The model then trained for seven more epochs with frozen embedding layers before failing to decrease test set loss for the next four epochs. The loss on the test set decreased substantially in the first epoch after freezing the embedding layers and decreased marginally on average over the next six epochs. While we may suppose that freezing the embedding layers prevented those layers from overfitting, this supposition cannot explain the simultaneous drop in train set loss. The value of the loss function on the train set decreased at a greater rate after freezing the embedding layers which is counterintuitive since since freezing reduced the parameter space, inhibiting fitting. Previous research has shown that freezing layers can reduce training time with minimal increase in the loss, but we are unaware of any research showing that freezing layers can decrease loss. We are curious for an explanation for this counterintuitive result.[3]

An appropriate performance metric must address right-censorship of the outcome. For a large share of observations, we do not observe the number of future years served but a lower bound on that value (due to censoring).[3] While we require a differentiable loss function to train the model, we use a non-differentiable metric, Harrell's concordance index c, to measure performance. The c index has a more direct interpretation for retention than the loss function.

[3]Typical metrics, such as mean squared error, require the observation of all outcome values over which the metric is computed.

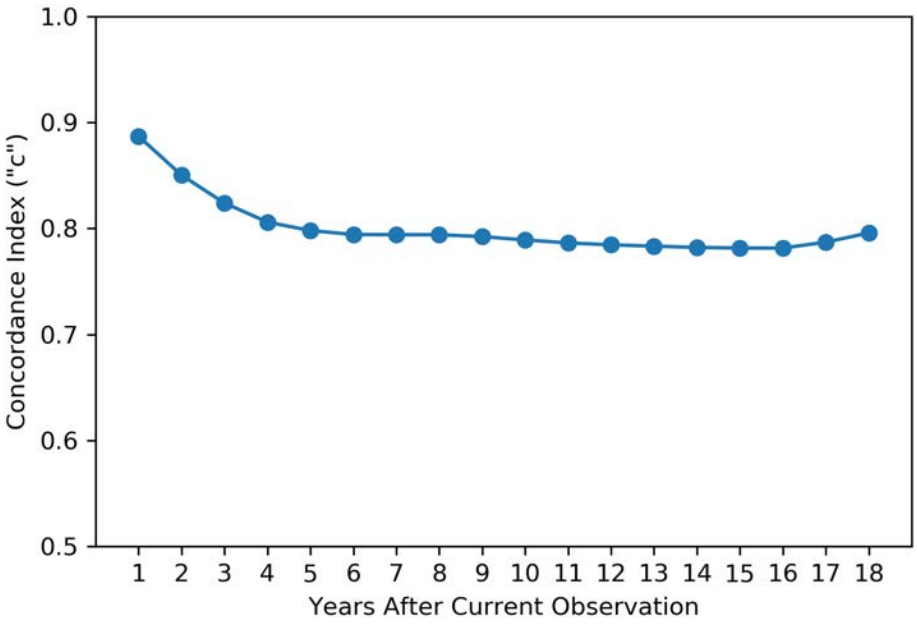

Figure 2: C-index on test set for each prediction time horizon

The c index is measured over the RMST predicted for each test set observation.[13] RMST is the expected value of the outcome over the maximum time interval of the predictions (in this case, 18 years). RMST is restricted because the model is estimated on data spanning only 19 years, so that the maximum possible observed value of the outcome is 18. If there was no censoring, then the estimate would be the mean number of years served until departure from service. The RMST is an under-estimate of the mean. Among all pairs of observations for which the smaller outcome value is uncensored, c measures the share of such pairs for which the observation with the smaller outcome value has a smaller predicted RMST.[14] Pairs for which the smaller outcome value is right-censored do not permit comparison of the outcome values so are not used to compute c. A c value of 0.5 means the model is producing uninformative predictions and a c value of 1.0 means the model is producing perfect orderings of predictions. Our model achieves a c of 0.725 on the test set.

We can compute c not only over RMSTs but also over probabilities of serving t consecutive additional years. To do so, we must discard observations fewer than t years prior to the most recent year (2018). Figure 2 plots c on the test set for each annual time horizon. c is 0.887 for the 1-year time horizon and drops to 0.802 for the 4-year time horizon, but does not drop much further for greater time horizons. c exceeds 0.78 for each time horizon. In this application, c is equivalent to the area under the receiver operating characteristic curve (AUROC).

To check that our model captures aggregate trends, we compare the actual survival curve for the entire test set with the survival curve predicted by the model. Each point on the predicted survival curve is the mean of the predicted survival probabilities over all test set observations for the given time horizon. As desired, Figure 3 shows that the model faithfully reproduces the actual survival curve.

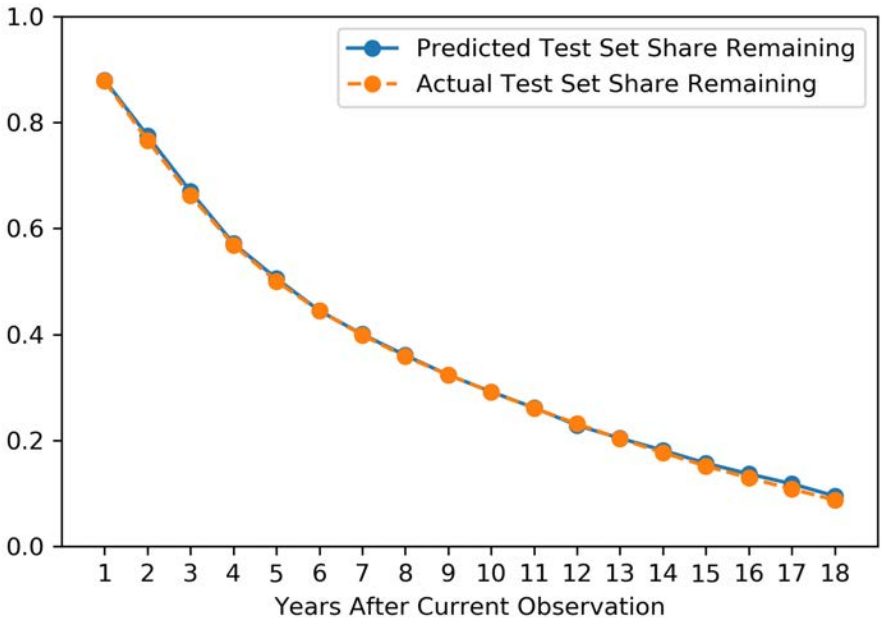

Figure 3: Predicted and actual aggregate survival probabilities

5. Conclusion

Using rich administrative data we began development of the RPM, a tool that can inform DoD leadership about anticipated retention at the service member level. The RPM is currently an application of a feed-forward neural network on a novel dataset which yields a c no worse than 0.78 for any given time horizon and 0.73 for RMST, on an out-of-sample subset of the data.

With further development, the RPM could be used within a larger econometric framework to provide counterfactual predictions at the service member level.[15] Future work can also identify features that have strong predictive power.[18] Additionally, this dataset can be used for methodological comparisons of various machine learning methods. Lastly, the loss decreasing after freezing is perplexing and warrants additional investigation.

References

[1] Our story, 2019. https://www.defense.gov/Our-Story/.

[2] Beth J Asch, James Hosek, Jennifer Kavanagh, and Michael G Mattock. Retention, incentives, and DoD experience under the 40-year military pay table. Technical report, RAND Corporation, 2016.

[3] Andrew Brock, Theodore Lim, James M Ritchie, and Nick Weston. Freezeout: Accelerate training by progressively freezing layers. *arXiv preprint arXiv:1706.04983*, 2017.

[4] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.

[5] Francois Chollet. *Deep learning with Python*. Manning Publications, 2018.

[6] Chin-Yuan Fan, Pei-Shu Fan, Te-Yi Chan, and Shu-Hao Chang. Using hybrid data mining and machine learning clustering analysis to predict the turnover rate for technology professionals. *Expert Systems with Applications*, 39(10):8844 – 8851, 2012.

[7] Richard L. Fernandez, Glenn A. Gotz, and Robert M. Bell. The dynamic retention model. Technical report, RAND Corporation, 1985.

[8] Dean A. Follmann, Matthew S. Goldberg, and Laurie May. Personal characteristics, unemployment insurance, and the duration of unemployment. *Journal of Econometrics*, 45(3):351 – 366, 1990.

[9] Michael F Gensheimer and Balasubramanian Narasimhan. A scalable discrete-time survival model for neural networks. *PeerJ*, 7:e6257, 2019.

[10] Ian J Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. Maxout networks. *arXiv preprint arXiv:1302.4389*, 2013.

[11] Glenn A Gotz, , and John McCall. A dynamic retention model for air force officers: Theory and estimates. Technical report, RAND Corporation, 1984.

[12] Claudia Grisales. Military recruitment, retention challenges remain, service chiefs say. *Stars and Stripes*, 2019.

[13] Frank E Harrell, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati. Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546, 1982.

[14] Frank E Harrell Jr, Kerry L Lee, and Daniel B Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361–387, 1996.

[15] Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep iv: A flexible approach for counterfactual prediction. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 1414–1423. JMLR.org, 2017.

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] David Knapp, Beth J. Asch, Michael G. Mattock, and James Hosek. An enhanced capability to model how compensation policy affects U.S. Department of Defense civil service retention and cost. Technical report, RAND Corporation, 2016.

[18] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017.

[19] Michael G. Mattock, Beth J. Asch, James Hosek, and Michael Boito. The relative cost-effectiveness of retaining versus accessing air force pilots. Technical report, RAND Corporation, 2019.

[20] Vishnuprasad Nagadevara and Vasanthi Srinivasan. Early prediction of employee attrition in software companies–application of data mining techniques. *Research and Practice in Human Resource Management*, 16:2020–2032, 2008.

[21] Rohit Punnoose and Pankaj Ajit. Prediction of employee turnover in organizations using machine learning algorithms. *International Journal of Advanced Research in Artificial Intelligence*, 5(9):22–26, 2016.

[22] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

[23] Eric Rosenbaum. IBM artificial intelligence can predict with 95% accuracy which workers are about to quit their jobs. *CNBC*, 2019.

[24] V. Vijaya Saradhi and Girish Keshav Palshikar. Employee churn prediction. *Expert Systems with Applications*, 38(3):1999 – 2006, 2011.

[25] Stefan Strohmeier and Franca Piazza. Domain driven data mining in human resource management: A review of current research. *Expert Systems with Applications*, 40(7):2410 – 2420, 2013.

[26] Lauren C. Williams. DOD can't hold onto cyber warriors. *GCN*, 2019.

[27] Yue Zhao, Maciej K. Hryniewicki, Francesca Cheng, Boyang Fu, and Xiaoyu Zhu. Employee turnover prediction with machine learning: A reliable approach. In Kohei Arai, Supriya Kapoor, and Rahul Bhatia, editors, *Intelligent Systems and Applications*, pages 737–758, Cham, 2019. Springer International Publishing.