Actuaries Summit
Think Differently

21-23 May 2017
Grand Hyatt Melbourne

Actuaries Institute

Stay or Go? The Science of Departures from Superannuation Funds

Prepared by Nathan Bonarius and Richard Dunn

Presented to the Actuaries Institute
Actuaries Summit
21 – 23 May 2017
Melbourne

This paper has been prepared for the Actuaries Institute 2017 Actuaries Summit.
The Institute's Council wishes it to be understood that opinions put forward herein are not necessarily those of the Institute and the Council is not responsible for those opinions.

© Nathan Bonarius and Richard Dunn

The Institute will ensure that all reproductions of the paper acknowledge the author(s) and include the above copyright statement.

Institute of Actuaries of Australia
ABN 69 000 423 656
Level 2, 50 Carrington Street, Sydney NSW Australia 2000
t +61 (0) 2 9239 6100 **f** +61 (0) 2 9239 6170
e actuaries@actuaries.asn.au **w** www.actuaries.asn.au

RICEWARNER

Insight like no other

Stay or Go? The science of departures from superannuation funds

Actuaries Summit 2017

22 May 2017

SYDNEY
Level 1
2 Martin Place
Sydney NSW 2000
P +61 2 9293 3700
F +61 2 9233 5847

MELBOURNE
Level 20
303 Collins Street
Melbourne VIC 3000
P +61 3 8621 4100
F +61 3 8621 4111

ABN 35 003 186 883
AFSL 239 191

www.ricewarner.com

Table of Contents

1. Executive Summary

1.1 Introduction

Superannuation funds operate in a competitive environment. Statistics from the Australian Prudential Regulation Authority (APRA) indicate that 44% of Responsible Superannuation Entities (RSEs) experienced negative cashflow (excluding investment returns) in the year to 30 June 2016. Consequently, funds are interested in developing retention strategies to prevent members leaving the fund to a competitor. In this paper, we analyse a unique data set from Rice Warner's Superannuation Insights study.

1.2 Method

Considering the Super Insights data set this paper aims to identify tools which may provide insight into the members who present the highest risk of exit. Specifically, the paper considers a range of both traditional and modern models, namely:

- Generalised Linear Models (GLM)
- Support Vector Machines (SVM)
- Random Partitions (RP)
- Random Forests (RF)
- Extreme Gradient Boosting (XGB)
- Ensemble models

1.3 Results

Overall we have been able to achieve a level of accuracy above our baseline 'naïve' estimator. Table 1 summarises the results from this analysis for the models considered.

	Basic	GLM	Random Partition	XG Boost	SVM	Random Forest	Ensembling - PPV	Ensembling - Accuracy
Total Accuracy	86.1%	86.6%	87.7%	89.5%	90.4%	92.4%	92.6%	86.6%
Positive Predicted Value	7.5%	10.8%	16.1%	30.2%	7.6%	47.7%	52.0%	24.2%
True Positive Rate	7.5%	10.8%	15.1%	30.2%	2.5%	22.5%	20.8%	37.1%

Overall these results reflect that:

- Non-traditional models (such as Random Forests) can be a viable, if not superior alternative to traditional models (such as Generalised Linear Models).
- Model error is unavoidable due to the low probability and complex nature of exit behaviours.
- Models can be trained to trade off on Type 1 or Type 2 errors. For example, funds could target a high level of true positive prediction to the detriment of other evaluation metrics.
- Models such as this can potentially be used to:

- Develop campaigns to target high-risk members and improve the return on investment for marketing campaigns.
- Drive improved understanding and appreciation of fund strengths and weaknesses through understanding the factors which identify members who are likely to leave.

This report was prepared and peer reviewed for the Actuaries Summit 2017, by the following parties:

Prepared by

Nathan Bonarius
Consultant
Telephone: (02) 9293 3722
nathan.bonarius@ricewarner.com

Peer Reviewed by

Michael J. Rice

Michael Rice
CEO
Telephone: (02) 9293 3704
michael.rice@ricewarner.com

Richard

Richard Dunn
Actuarial Analyst
Telephone: (02) 9293 3713
richard.dunn@ricewarner.com

22 May 2017

2. Overview

2.1 Problem statement

Superannuation funds routinely lose a portion of their membership each year. Membership movements can be driven by a number of factors including rollovers to competitors, member establishment of their own self-managed fund, transition of lost accounts to the Australian Tax Office (ATO) or the transfer of assets on meeting a condition of release.

Statistics from the Australian Prudential Regulatory Authority (APRA) indicate that 44% of Responsible Superannuation Entities (RSEs) experienced negative cashflow (excluding investment returns) in the year to 30 June 2016. Member retention is essential to stemming this outflow. In this paper we explore models which can be used to assist funds to predict member exits, using a large sample of data from Rice Warner's annual Superannuation Insights study which represents over 14.4 million member accounts with 1 million member exits over a period of 3 years.

2.2 A note on notation

Throughout this paper we use the following notation:

- E_i denotes the status of member "*i*"; 1 for exited in the year in question, 0 for remain within the fund
- $\widehat{E}_i$ to denote an estimate of the value of E_i for member "*i*"
- F_i as the full set of available information regarding account "*i*" at the time of estimation

2.3 Model baseline

Throughout this report we have benchmarked our predictions against a simple empirical estimator. While wide ranging, estimators of this type typically leverage the average rate of exit as a constant estimator of the form (where notation is defined as in section 2.2):

$$\Pr\left(\widehat{E}_i = 1 \,\middle|\, F_i\right) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}_{\{E_j=1\}}$$

Estimators of this form practically constitute a random "guess" with respect to the mean and thus perform poorly.

In constructing predictive models actuarial practitioners have traditionally turned to generalised linear models for their relative simplicity, robust statistical basis and ability to tackle a number of predictive problems. However, with the increasing volume of data available models of this sort are being challenged by newer models from the machine learning community which are more computationally intensive. In this paper we contrast both traditional actuarial and new approaches to this problem.

3. Data analysis

3.1 The sample

Our sample is taken from Rice Warner's flagship research project *Superannuation Insights*. This project includes anonymised individual member records for 14.4 million[1] distinct superannuation accounts representing over 20 superannuation funds (including retail, industry, public sector and corporate funds) across a set of 42 categorical and numeric fields over three years.

In aggregate these fields span the superannuation experience and allow membership classification across demographics, investment preference, insurance coverage as well as annualised transactional data. Appendix A constitutes a complete listing of the fields included within the database used in this paper though not all collected fields were made available for exited accounts.

We note, predicting exits is difficult in that it is a low probability event. Exits in the considered sub-sample have historically occurred in approximately 7.5% of available accounts and this excludes partial withdrawals.

3.2 Demographics

Superannuation membership is diverse covering most of the Australian population. Engagement with superannuation is suspected to be strongly linked with a number of demographical indicators such as age, account balance, gender, occupation and socio-economic factors.

The main variables of interest from our sample are:

- age (calculated from date of birth)
- gender
- tenure in the fund

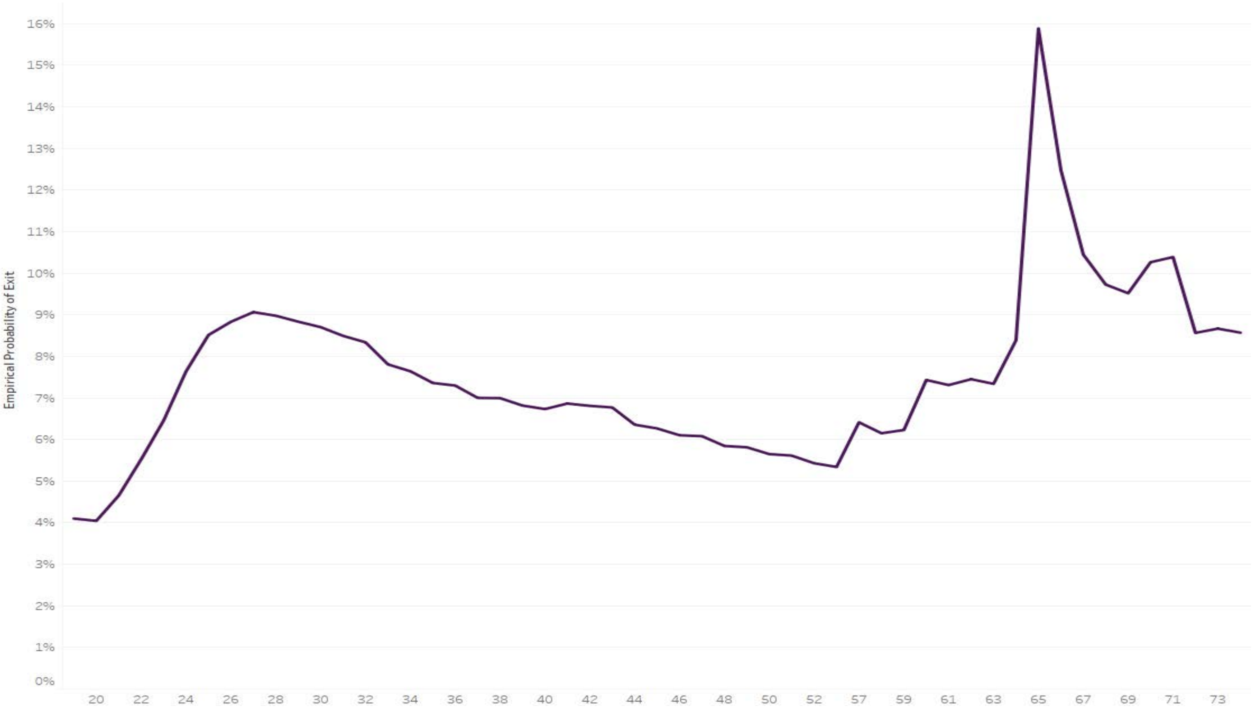
3.2.1 Age

Graph 1 shows the probability of exit as a function of age in our sample.

The graph shows results which are consistent with expectations from behavioural finance. It demonstrates that individuals place far greater emphasis on decisions which influence short term outcomes rather than longer dated ones. Once members have selected or defaulted into their superannuation fund early in their careers, they then tend to disengage from taking an active interest in their superannuation account. As a result, member exits originally peak around age 30 and remain subdued until members begin to plan for retirement from age 50.

[1] Relative to the total market of 29.674 million accounts (Rice Warner's *Superannuation Market Projections Report 2017*)

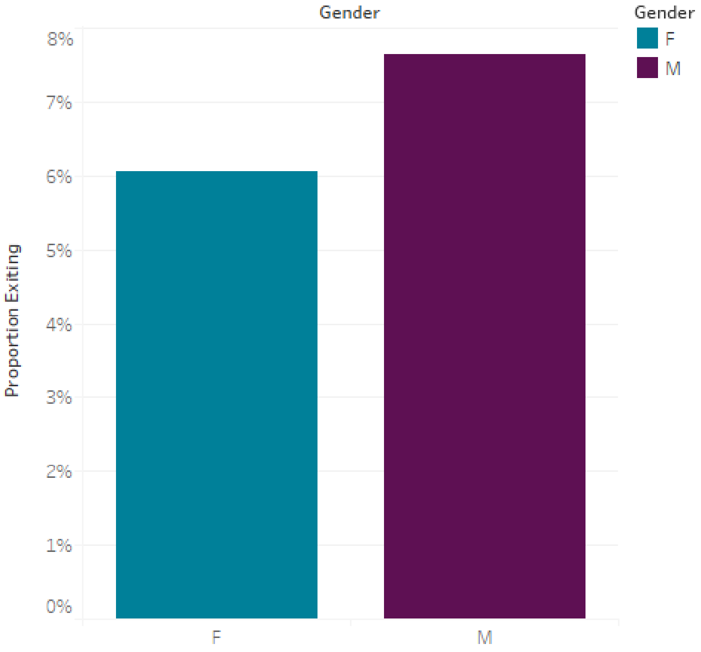
Graph 1. Empirical probability of exit as a function of age



3.2.2 Gender

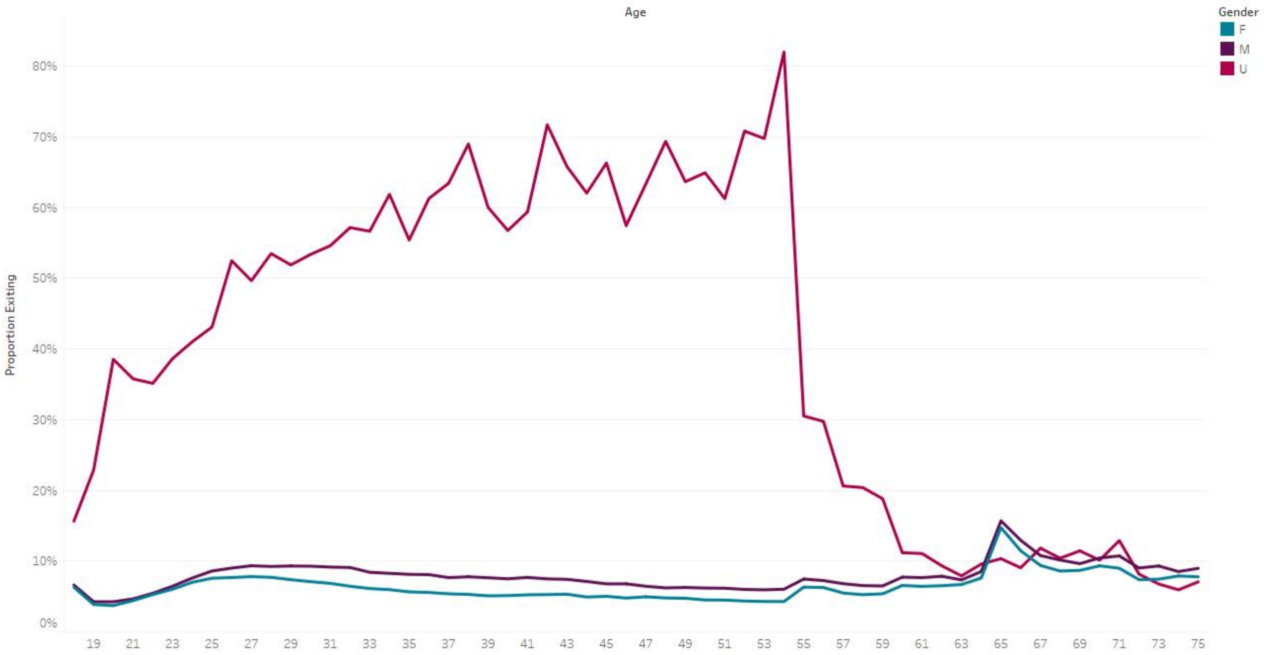
Graph 2 shows men are, on average, 26% more likely to leave a fund than a female. In addition, Graph 3 shows that if a fund does not know the gender of their client the member is far more likely to exit the fund at younger ages[2].

Graph 2. Empirical probability of exit as a function of gender



[2] Noting that while still large in absolute terms, the sample size of "Unknown" genders is comparatively small (approximately 0.3% of the sample).

Graph 3. Empirical probability of exit as a function of gender and age

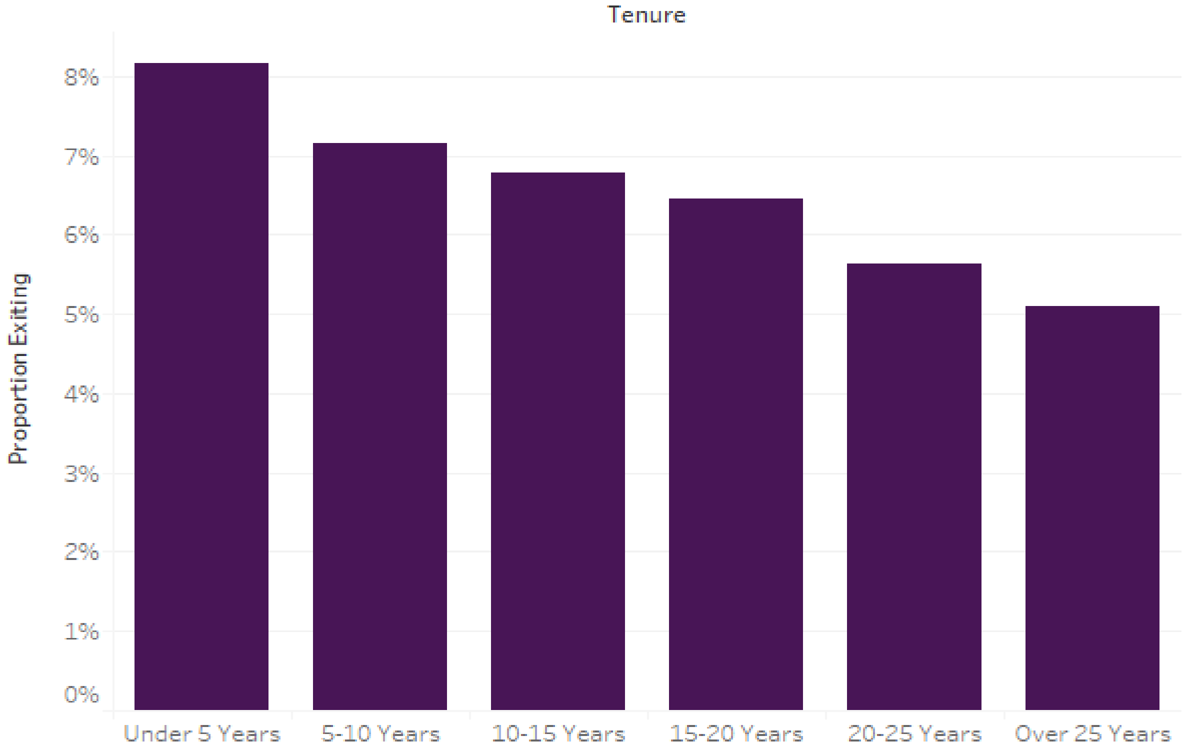


3.2.3 *Tenure*

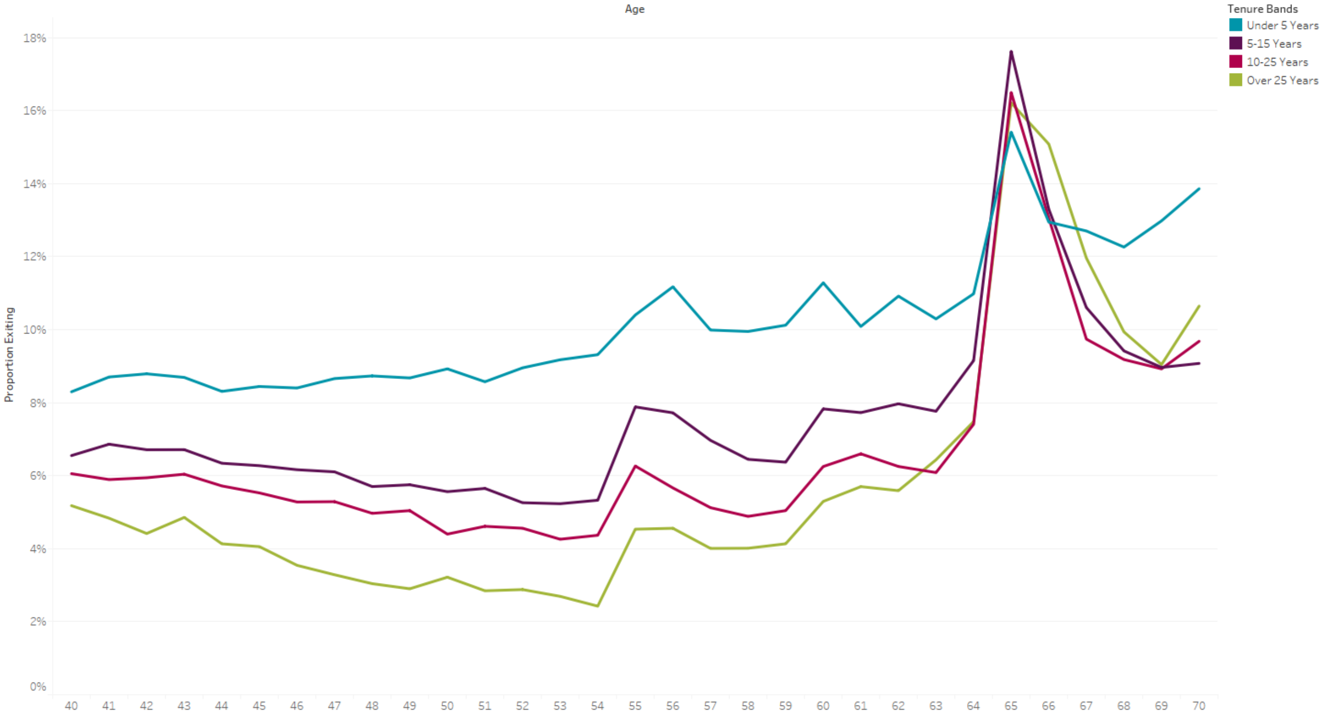
Fund tenure is defined as the time in years since a member joined the fund and reflects, (after accounting for age) the level of loyalty (or perhaps disengagement) a member has with the fund.

Graph 4 shows the rate of exit is lower for longer serving members. This could demonstrate that members who have stayed in the same fund for a longer time horizon are comparatively more likely to be satisfied with their fund performance and therefore less likely to consider an external transfer. Graph 5 shows that the relationship holds true even when controlling for age, validating the idea that tenure is a useful predictor.

Graph 4. Empirical probability of exit as a function of tenure in the fund.



Graph 5. Empirical probability of exit as a function of tenure in the fund and age.



3.3 Transaction data

We have considered the following variables from financial transactions in our estimates:

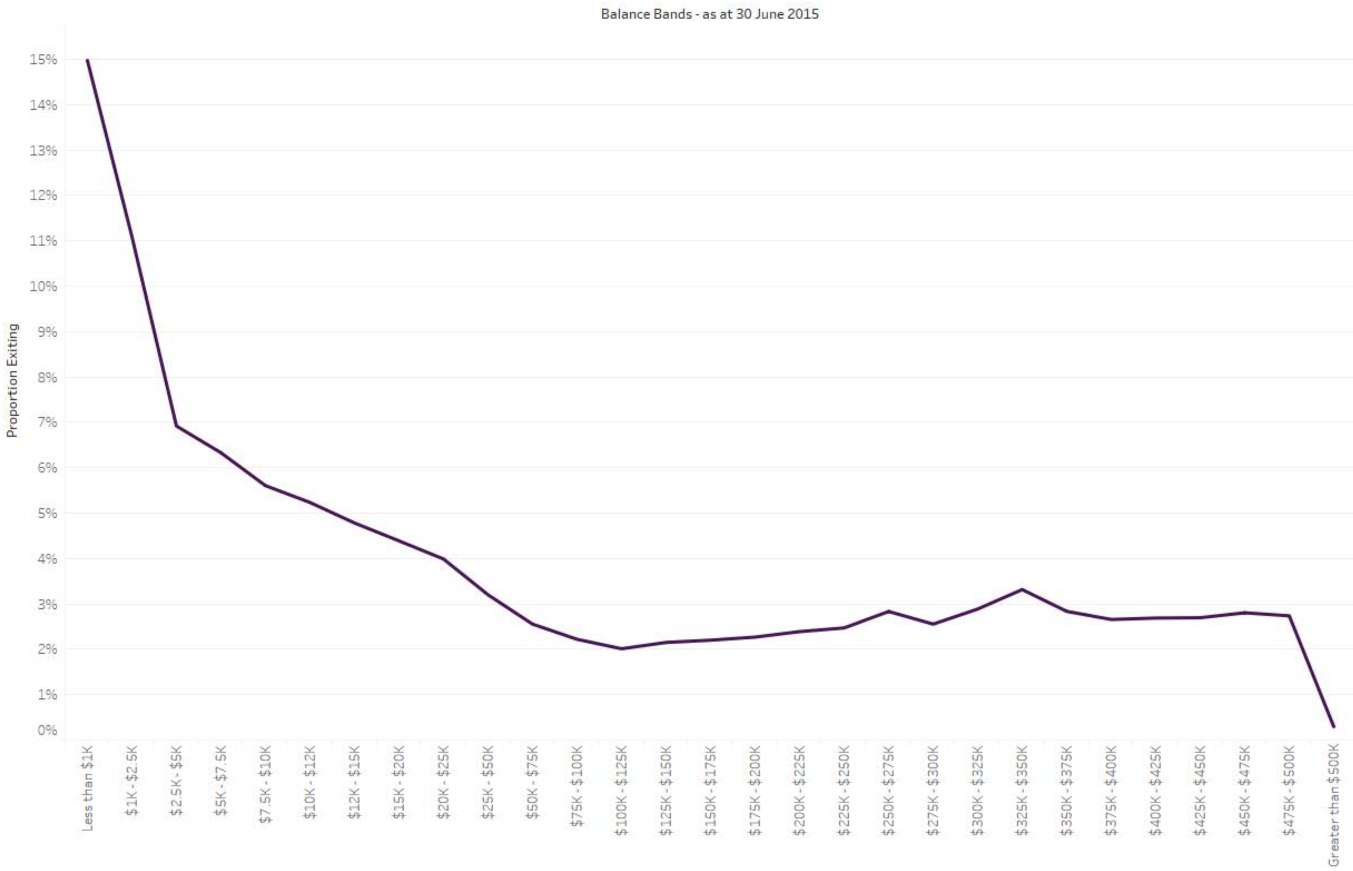
- account balance; both in the current year as well as its movement over the past financial years, and
- voluntary member contributions

3.3.1 Balance

Graph 6 demonstrates that smaller account balances are more likely to be represented in exits from the fund. This could follow from:

- Members with low balances being more likely to consolidate small balances to their primary (or larger) account.
- Higher levels of exits at younger ages (where members have not yet accrued large balances).
- Active but disengaged members accruing higher balances but being comparatively less likely to exit relative to inactive members due to being not subject to automatic processes such as ATO Lost Super transfers.

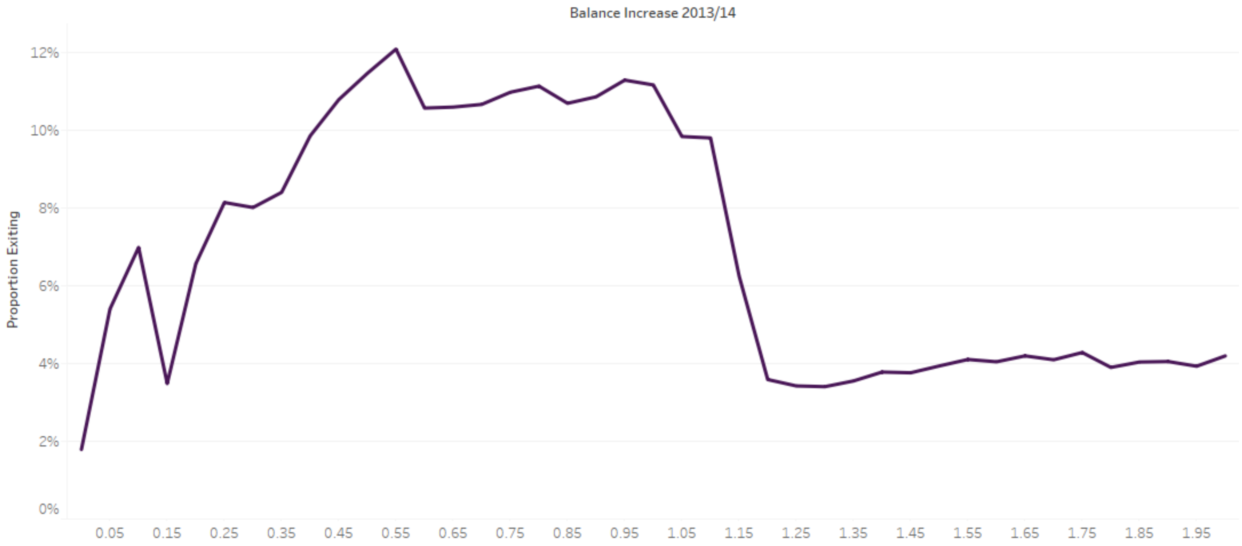
Graph 6. Empirical probability of exit as a function of balance



Graph 7 demonstrates that account balances that are declining are more likely to exit the fund. This may be a result of several factors:

- Members making partial rollovers in years before exiting
- Members leaving funds with poor investment performance
- Members consolidating inactive small accounts where fees and premiums outweigh investment returns

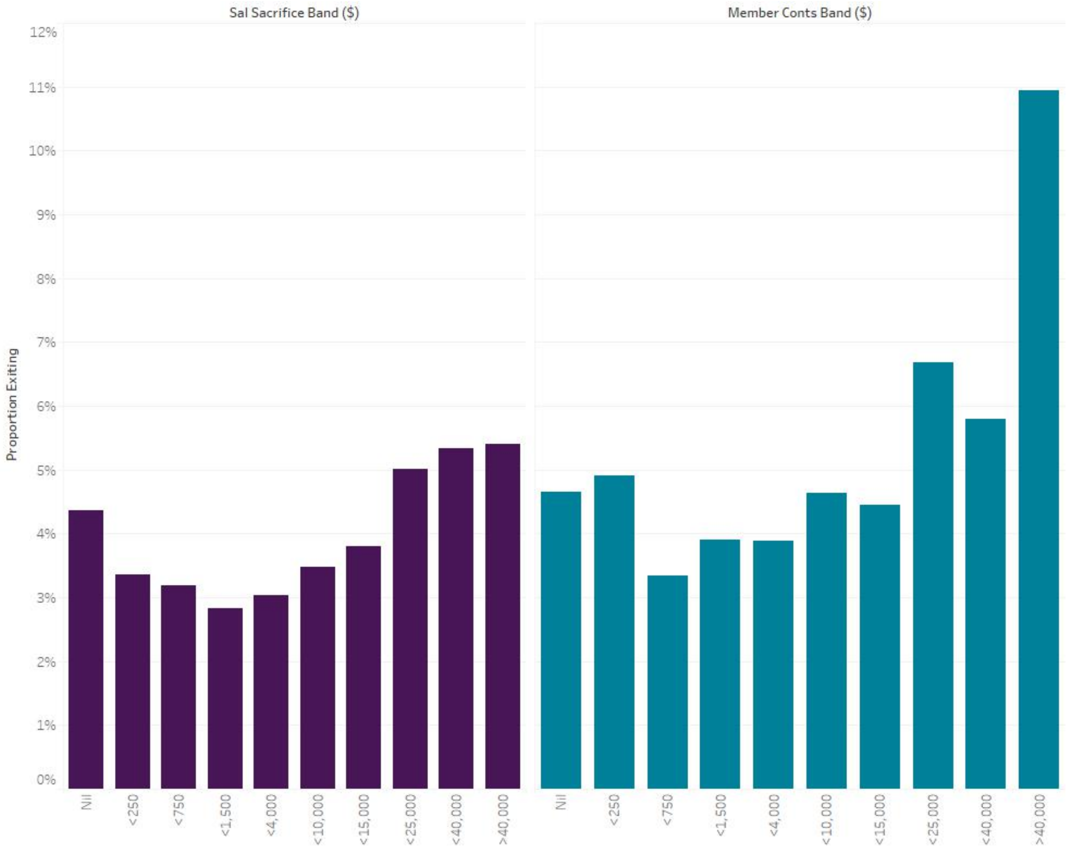
Graph 7. Empirical probability of exit as a function of balance relative to the preceding year



3.3.2 Contributions

Contribution behaviour can also provide insight into a member's level of engagement. Graph 8 shows, except for members who make nil or minor contributions, as contributions – both concessional and no concessional – rise, so too does the probability of a member exiting the fund. However, members making small contributions are less likely to exit the fund than those who make none at all.

Graph 8. Empirical probability of exit as a function of voluntary contributions



This is likely a result of members who make contributions above the mandatory super guarantee being on average more financially aware than their non-contributing counterparts. Thus, these members are more readily able and willing to assess fund performance and identify if a fund change is needed. Further, this effect is compounded in the case of members who have a sufficiently high salary to make non-concessional member contributions.

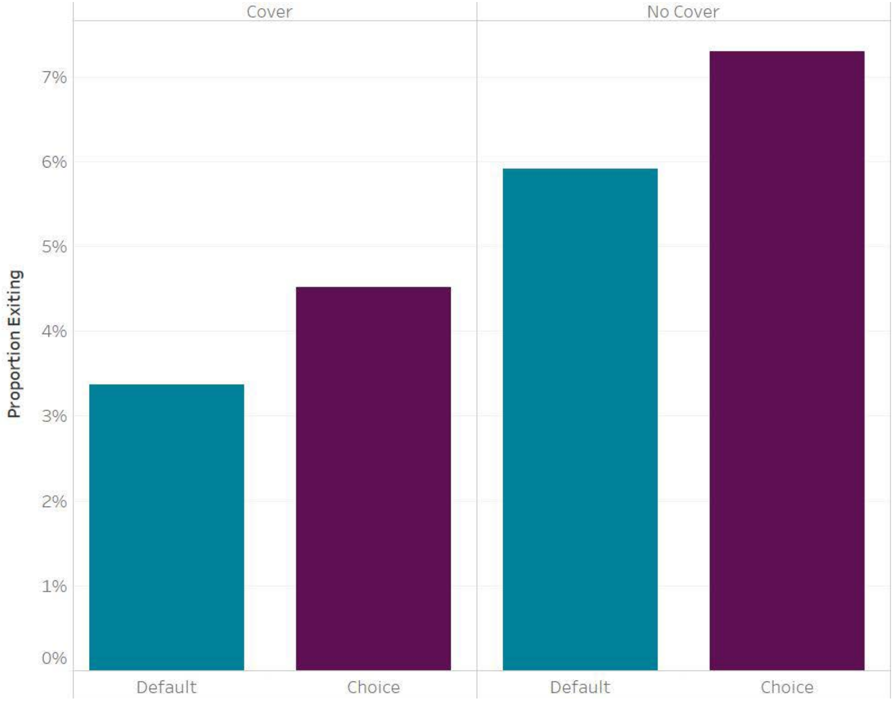
3.4 Investments choice and opt out of insurance

As noted above, some forms of engagement and involvement with a superannuation fund may act as a leading indicator of a member's intention to exit.

Graph 9 shows the probability of member exit by those who opt out of default insurance and those who make an investment choice. The result is consistent with the result from voluntary contributions in that:

- Members who adopt the default option are comparatively more likely to remain in the fund than members who make a personal investment choice.
- Members who have opted out of insurance cover are less likely to remain in their current fund relative to members who still have cover within their Superannuation.

Graph 9. Empirical probability of exit as a function of investment and insurance choices



3.5 A note on inter-fund heterogeneity

It is worth noting that there is significant heterogeneity in the rate of exit between different funds. This factor will capture a number of variables, including:

- Demographic differences in the membership basis, such as some of those already mentioned but also other unknowns in our sample including occupation, employer/industry, take up of advice etc.
- Internal fund actions e.g. retention campaigns, branding and advertising.

Thus, in the interest of providing unbiased analysis a categorical predictor for the fund has been considered in the model building process.

4. Methodology

4.1 Predictors

Following the analysis in Section 3, we identified the set of predictors we consider most useful to predict future exits. In the interest of parsimony, the universe of predictors has been restricted based on comparative predictive power on both a standalone and an integrated (i.e. including interactions) basis. Considering this, the predictors considered include:

Table 1. Predictors

Demographics	Transactional Data	Other
Age	Balance in 2015	Insurance Cover Status
Gender	Balance in 2014	Investment Choices
Tenure	Balance in 2013	
	Member Contributions	
	Salary Sacrifice Contributions	

4.2 Models considered

We have modelled the propensity to exit using the following models. We have evaluated these models against the basic estimator defined in Section 2.2:

- Generalised Linear Models (GLM).
- Support Vector Machines(SVM)
- Random Partitions (RP)
- Random Forests (RF)
- Extreme Gradient Boosting (XGB)
- Ensemble models

All modelling was undertaken in the statistical package R.

4.2.1 Generalised linear model (GLM)

Generalised linear models have formed the backbone of the statistical modellers toolkit since they gained widespread popularity in the 1980s for their ability to model bounded response variables with non-normal errors. In the context of exit modelling this is a highly useful as whether a member exits from a fund is a binary response outcome. Consequently, we have used a GLM with a logit-link function of the form:

$$\Pr\left(\hat{E}_i = 1 \mid F_i = \boldsymbol{x}\right) = \frac{e^{-\boldsymbol{ax}}}{1 - e^{-\boldsymbol{ax}}}$$

Where $\boldsymbol{ax}$ is some linear function of the predictors with degree greater than or equal to 1.

4.2.2 Support vector machine (SVM)

Support Vector Machines (SVM) are a sub-class of machine learning algorithm which are typically used for models where the response-predictor relationship of interest is either categorical, quantitative or a combination of both. In line with this when fitting an SVM model, the multi-dimensional response-predictor relationship is translated to a series of points in n-dimensional space (where n is number of features considered) with the value of each feature being the value of a coordinate. Using this n-dimensional space the model fits a hyperplane from which the fit can be used to make predictions and classifications.

4.2.3 Random Partition models (RP)

Random partitions are used when global ordering is required. It partitions the data with respect to a pre-defined set of exclusive and continuous ranges that cover the entire domain of the partition key and from this makes forecasts.

4.2.4 Random Forest models (RF)

Random forests work as a collection of randomized decision trees. In general, by building a large set of decorrelated decision trees a modeller can simulate a large set of potential classifications for a forecasted record. Given these potential classifications and the fact that these simulations are decorrelated the law of large numbers is invoked to provide an estimate of the statistically most likely classification as per:

$$\Pr\left(\hat{E}_i = 1 \mid F_i = \boldsymbol{x}\right) = E_{E_i}[\Pr(E_i = 1)|F_i = \boldsymbol{x}]$$

Where E_{E_i} denotes expectation with respect to the random parameter, conditionally on the complete data set inclusive in $\boldsymbol{x}$.

4.2.5 Extreme Gradient Boosting models (XG Boost)

Gradient boosting works to produce estimates for regression and classification problems through the production of a large number of weak prediction models (such as lightly correlated decision trees). Armed with these models the boosting algorithm then combines them to assemble a model which minimises the bias and variance of the estimate with respect to the data.

4.2.6 Ensemble modelling

Ensemble models refer to any model which is a combination of two or more separate “sub models”. In general, the logic underlying models of this kind is that the error terms inherent to the sub models will either diversify or mitigate one another (through the errors being negatively correlated) to together produce a more accurate model.

4.3 Evaluation

We have considered the following evaluation metrics:

- Log-Loss
- Improvement on a basic estimator
- Confusion matrices

4.3.1 Log-Loss

Log-loss compares the predictive quality of an estimator after accounting for the relative likelihood of events occurring. Defined mathematically for an event $\hat{E}_{i,model}$ with corresponding probability p as:

$$L = \sum_{i=1}^{n} -\hat{E}_{i,model}\log(p_i) - \left(1 - \hat{E}_{i,model}\right)\log(1 - p_i)$$

In practical terms the log-loss of a model effectively assesses the probability-weighted sample accuracy, with lower values of the metric denoting improved accuracy.

The log-loss imposes a harsher penalty for incorrect predictions with high levels of confidence. In practice, given the metric will become infinitely large for values of p = 0 or 1, these will be replaced with a 1E10^-15 or 1-1E10^-15 respectively.

4.3.2 Improvement on a basic estimator

Crucial for any estimator to be deemed successful is that it adds value over and above an uninformed "naïve" approach. Consequent to this, and considering the "basic" estimator identified in Section 2.2 we consider an "improvement" metric of the form:

$$M_{imp} = \frac{1 - \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{\hat{E}_{i,model}=E_i\}}}{1 - \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{\hat{E}_{i,\text{naïve}}=E_i\}}} - 1$$

Overall, M_{imp} quantifies the number of correct predictions made by the model under consideration relative to the theoretical number of correct predictions made by the naïve estimator[3]. On a practical basis, this metric can therefore be translated as the percentage increase (or decrease) in accuracy garnered by implementing a more complex model.

4.3.3 Confusion Matrices

Confusion matrices present a simple way of displaying the distribution of predictions. Specifically, the matrix splits out predictions into correctly predicted positives and negatives as well as respective type 1 (false positive) and 2 (false negative) errors. This display is useful as it allows a modeller to identify the regions in which their model performs well and the areas in which the model needs improvement.

The confusion matrix also provides number of important derivative metrics. Specifically, these include:

- total accuracy
- positive predictive value (the proportion of predicted exits which did exit)
- true positive rate (proportion of exits which were correctly predicted)

[3] Approximately 86%.

5. Results

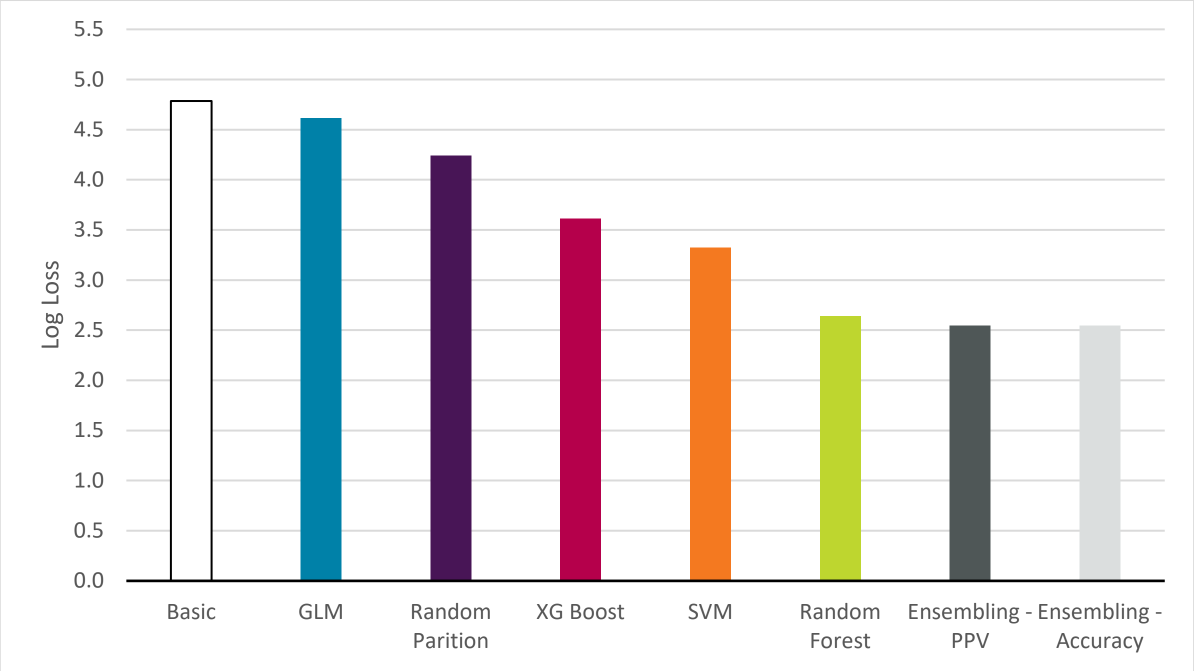
Considering these models and metrics, we now turn to the results produced. This section considers a sub-sample of the Super Insights data comprising of a training set of 5,009,004 and test set of 556,557 distinct member accounts which were at risk of leaving over the 2014-2015 financial year. Appendix B provides summary statistics with respect to the sample.

5.1 Log-Loss

As was previously noted, the log-loss effectively measures the probability-weighted sample accuracy, with lower values of the metric denoting improved accuracy. Figure 9 contrasts the relative log-loss metrics of the considered models and reflects that:

- Each data driven model outperforms the naïve estimator, with the level of improvement varying significantly between models.
- In general, the random forest significantly outperforms its competitors with approximately half the log-loss of the naïve estimator.
- Despite its popularity the GLM approach provides very little marginal improvement over the naïve estimator. This is consistent even when the order of polynomial considered in the model is increased dramatically.

Graph 10. Log Loss function value for various metrics



5.2 Improvement on a basic estimator

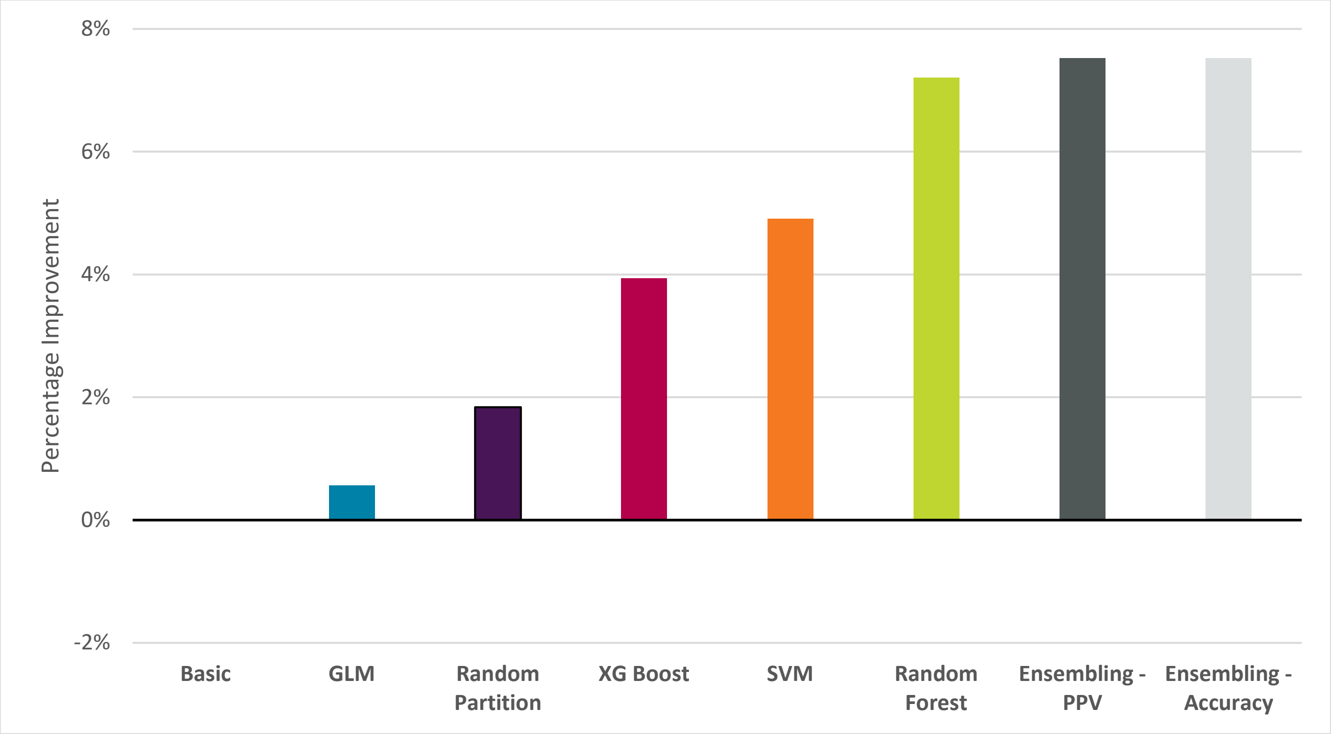
In line with the results for the log-loss – contrasting with the basic estimator reflects that:

- The GLM performs poorly, only providing a 0.57% improvement increase over the baseline naïve model.
- Ensembling the SVM, XG Boost and the Random Forest models to optimise total accuracy in a linearly optimal way yields results which provide an increase 13 times that of the standard GLM.

- Aside from ensembling, random forest models provide the best predictive power of the models under consideration.

It is worth noting at this point that while the absolute value of the percentage improvement is low (below 8%) the effect of even a small increase in accuracy provides a marked improvement in the number of member outcomes correctly predicted.

Graph 11. Percentage improvement on a naïve estimator



5.3 Confusion Matrices

Confusion matrices, unlike other metrics are not able to be 'optimised' in a sense. However, they have strength in that they allow identification of the relative strengths of a model in terms of both its type 1 and type 2 errors. Measurement of both errors allows the user to make model trade-offs where the errors have real world consequences. In the context of our work this would be the scenario of running a campaign to increase member retention and balancing the risks of contacting members who won't leave the fund this year (Type 1 errors) against the risk of not contacting members who will (Type 2 errors).

Table 2 reflects potential diversity in the different metrics. For example, although there was only an 8% increase in total accuracy from using the optimal ensemble model, this translates to nearly a 7-fold increase in the forecasting ability of the model in terms of the predicted exits (PPV). Further to this the table reflects:

- in using the optimal ensemble model the number of correct exit predictions can be increased to over 50%, and
- there is significant variability in the positive predictive power of the models considered.

Table 2. Selected Confusion matrix statistics for each model

	Basic	GLM	Random Partition	XG Boost	SVM	Random Forest	Ensembling - PPV	Ensembling - Accuracy
Total Accuracy	86.1%	86.6%	87.7%	89.5%	90.4%	92.4%	92.6%	86.6%
Positive Predicted Value	7.5%	10.8%	16.1%	30.2%	7.6%	47.7%	52.0%	24.2%
True Positive Rate	7.5%	10.8%	15.1%	30.2%	2.5%	22.5%	20.8%	37.1%

5.4 Importance Plots

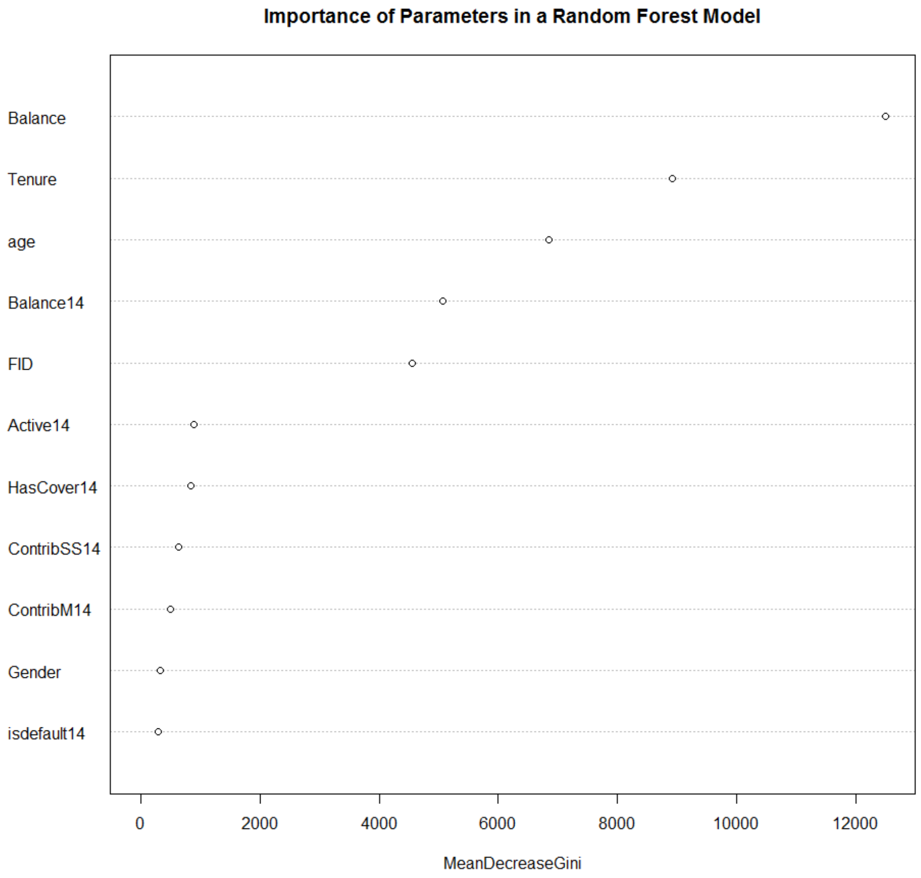
In the context of these models we have examined the relative contribution of the predictors to model performance. Graph 12 and Graph 13 visualise this importance in the case of both the Random Forest and XG boost models and demonstrate that the models is largely driven by a small handful of the same predictors in both cases, namely:

- Current account balance
- Tenure in the fund
- Member age
- Balance in the previous financial year
- Fund specific effects

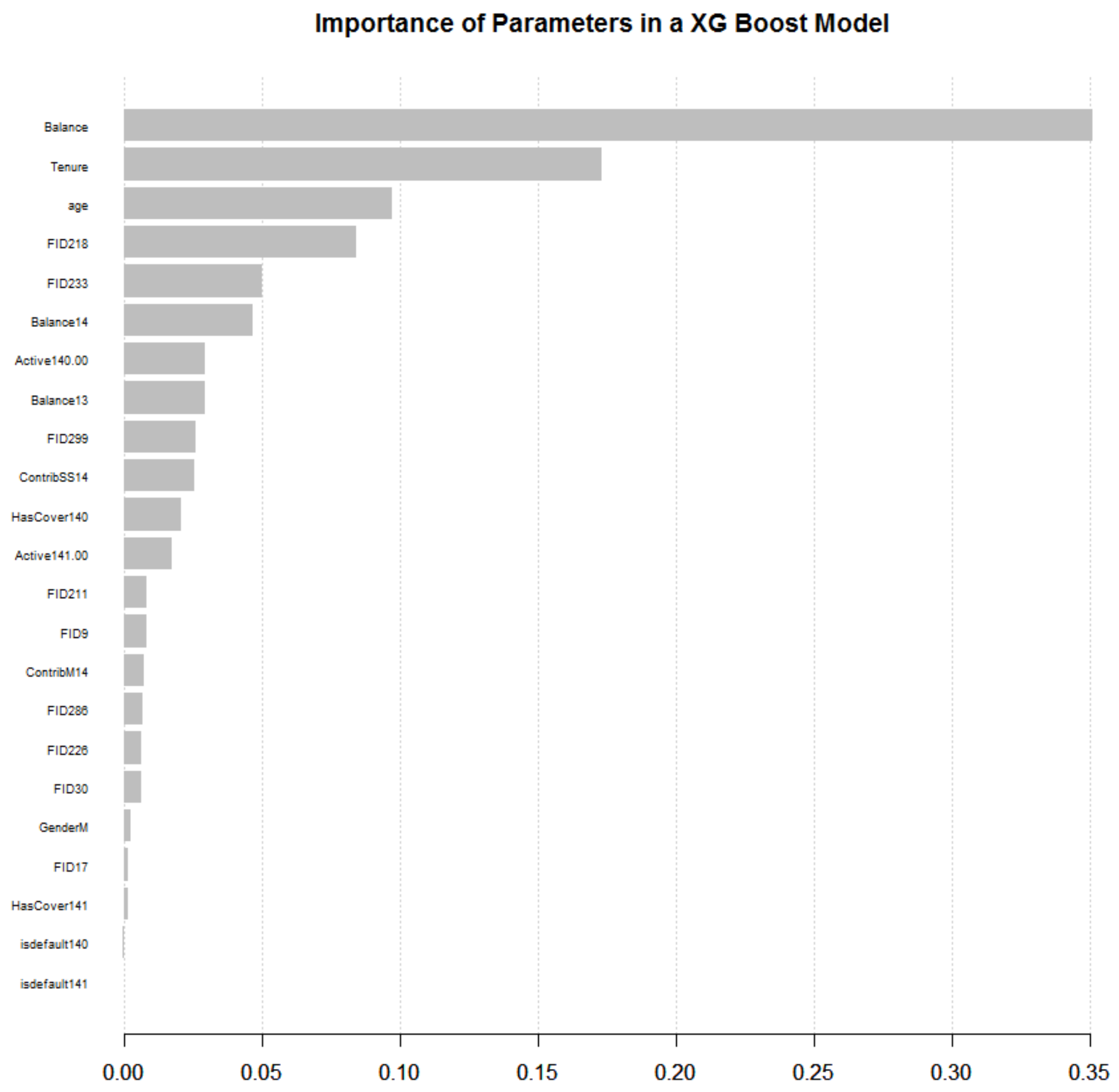
Further, it is interesting to note that:

- Gender is limited in its use as a predictor despite the comparative likelihood of men to exit relative to women. This may indicate that the gender effect reflects the influence of other factors such as balance and age.
- Member investment choice (or lack thereof) adds limiting predictive power. This may stem from the suitability of modern default options, the comparatively small number of choice members or the high incidence rates of null entries in the data induced by considering inter-year behaviour.

Graph 12. Importance plot – Random Forest model



Graph 13. Importance plot – XG Boost model



5.5 Conclusion

From the results we note the following:

- Non-traditional models (such as Random Forests) can be a viable, if not superior alternative to traditional models (such as Generalised Linear Models).
- Model error is unavoidable due to the low probability and complex nature of exit behaviours.
- Models can be trained to trade off on Type 1 or Type 2 errors. For example, funds could target a high level of true positive prediction to the detriment of other evaluation metrics.
- Models such as this can potentially be used to:
  - develop campaigns to target high-risk members and improve the return on investment for marketing campaigns

- drive improved understanding and appreciation of fund strengths and weaknesses through understanding the factors which identify members who are likely to leave

5.6 Further work

While every effort has been made to ensure a high degree of accuracy in the model it is important to concede that there are extensions and improvements that could be made:

- incorporating further predictors for the reason of exit – using indicators not available in our dataset
- exploring other models or selection of model parameters
- optimisation of models using other evaluation metrics targeted to a marketing campaign.

Appendix A Super Insights Data Fields

Table 3. Accumulation member information

Member Information
Product Name and/or Division
Sub-plan and/ or Employer
Member Number
Employer
Date joined fund
State
Postcode
Country
Date of birth
Gender
TFN Supplied
Financial Information
Opening Balance
Closing Balance
SG contributions
Salary sacrifice contributions
Co-Contributions
LISC Contributions
Personal Contributions
Spouse contributions
Transfers in
Interest/Income
Contribution tax
No TFN tax
Administrative fee
Other fees
Total fees
Lump sum withdrawals
Insurance Information
Beneficiary count
Default death cover
Default TPD cover
Voluntary death cover
Voluntary TPD cover

Salary continuance cover
Occupation classification
Benefit Period
Waiting Periods
Death cover premium
TPD cover premium
Salary continuance cover premium
Investment Information
Count of investments
Default investment option
Investment choice name
% of member assets in the investment or $ in the investment
Investment Switches
Number of switches
Date of switch
Option switched from
Option switched to
Advisory Services
Type of advice
Frequency
Cost of advice
Internal or external advice

Table 4. Exited member information

Exited members
Member Number
Date of Birth
Gender
State
Postcode
Date joined fund
Reason for exit
Payment amount
Date of exit
Full Exit
Remaining balance in fund
Rollover fund name
SMSF Flag

Exited members
Rollover fund Unique Superannuation Identifier (USI)
Rollover fund type (if available)
Rollover fund ABN
Rollover fund Superannuation Fund Number (SFN)

Table 5. Pensioner member information

Member Information
Product Name/Division
Account type
Member Number
Date started pension
Date of birth
State
Gender
Postcode
Country
TFN Supplied
Financial Information
Opening Balance
Closing Balance
Transfers in
Interest/Income
Administration fees
Other fees
Total fees
Lump sum withdrawals
Pension payments
Investments
Default investment options
Number of investment choices
Investment choice name
% of member assets or $ in the investment option