On the prediction of claim duration for income protection insurance policyholders

Qing Liu*,[1], David Pitt[2], Xueyuan Wu[1]

[1]Centre for Actuarial Studies, Faculty of Business and Economics, The University of Melbourne

[2]Department of Applied Finance and Actuarial Studies, Faculty of Business and Economics, Macquarie University

October 2012

Abstract

This paper explores how we can apply various modern data mining techniques to better understand Australian Income Protection Insurance (IPI). We provide a fast and objective method of scoring claims into different portfolios using available rating factors. Results from fitting several prediction models are compared based on not only the conventional loss prediction error function, but also a modified loss function. We demonstrate that the prediction power of all the data mining methods under consideration is clearly evident using a misclassification plot. We also point out that this predictability can be masked by looking at just the conventional prediction error function. We then suggest using principal component analysis to increase understanding of the rating factors that drive claim durations of insured lives. We also discuss and compare how different variable combining techniques can be used to weight available predicting variables. One interesting outcome we discover is that principal component analysis and the weighted combination prediction model together provide very consistent results on identifying the most significant variables for explaining claim durations.

Key words: Income Protection Insurance; data mining; principal component analysis; weighted combination.

1 Introduction

Data mining is the type of analysis made possible by modern computers, which uses powerful processors to mine through large databases to reveal previously unsuspected or unquantified trends and relationships. The manual extraction of patterns from data has existed for centuries. Some examples of identifying patterns in data include Bayes' theorem in the 1700s and regression

*Corresponding author. Centre for Actuarial Studies, Faculty of Business and Economics, The University of Melbourne, VIC 3010, Australia. Email: liuq@student.unimelb.edu.au.

analysis in the 1800s. The increasing power of computer technology has dramatically increased data collection, storage, and manipulation ability. As data sets have increased in scale and complexity, indirect and automated data processing methods such as neural networks, cluster analysis, and decision trees have become increasingly popular. In more recent years, actuaries have started to apply data mining methods. In early 2001, Senensky and Polon founded a consulting company called Claim Analytics with the objective of using data mining tools for predictive modelling to assist major insurance companies with claim scoring, pricing, reserving and fraud detection. This paper is inspired by a case study done by Senensky and Polon on prediction of return to work for group insurance claimants with long term disabilities in 2004. We further explore how other data mining techniques can be applied to predict duration classes of insured lives under Income Protection Insurance (IPI). Income Protection Insurance plays a significant role in maintaining the quality of life of individuals, of working age, who become unable to work due to a non-work related injury or an illness. It achieves this by providing such insured lives with a proportion of their usual salary during the time that they are unable to work. A major risk for providers of the IPI business comes from the fact that the claim duration can vary considerably for different IPI policies. Therefore a good understanding of the durations of IPI claims is an important input to actuaries' pricing and reserving calculations. We have obtained Australian IPI claim data from the Actuaries Institute of Australia (IAAust) income protection insurance policy database, which contains comprehensive information on policyholders who have purchased insurance from the main Australian providers of IPI. Data are recorded for each policyholder based on the information provided in the insurance proposal form. The data were previously analysed by Pitt (2007) and contain all claim records beginning in calendar year 1995. The total number of claim records in respect of calendar year 1995 is 8863. These claims were followed until termination or the end of calendar year 1998, whichever occurred first. Available information provided in the data set includes the duration of each policyholder's claim, the age of the claimant at the onset of disability, the definition of disability used in assessing whether the policyholder is eligible for a benefit under the policy, the gender of the policyholder, the occupation class of the policyholder (classified into four levels, see the Report of the IAAust Disability Committee, 1997), the rate of benefit payable monthly, the type of benefits payable (increasing in line with inflation or level), the smoker status of the insured life and the deferment period specified in the insurance contract. Table 1 provides a summary of the potential rating factors recorded for each of these claimants along with the coded SAS variable name and a brief description of the variable.

Table 1: IPI Data Fields

Variable	**Description**	**SAS code**
Duration	Duration of the claim (recorded in days). This is the number of days from when the sickness began until recovery (or censoring), less the deferment period.	durn2
Age	Age at the date of claim commencement	age
Terminate	An indicator of whether the claim was observed to terminate or was censored	terminate
Disability Definition	Own occupation for which the insured person is reasonably suited by education, training or experience, or any occupation after an initial period. (Indicator variable for any occupation after initial period)	poldesnew3
Sex	Indicator variable for gender; Male = 1.	sex1
Occupation	Class Occupation is grouped into four levels: A, B, C or D as described in IAAust Disability Reports	occupB, occupC, occupD
Benefit Rate	Monthly benefit rate in dollars	benrate
Benefit Type	Level or Increasing Benefits. (Indicator variable for increasing benefits)	bentypnew2
Medical Evidence	Medical Exam required or Automatic Acceptance. (Indicator for medical exam required)	medevid1
Contract Type	Level Premiums or Stepped Premiums. (Indicator variable for Level Premiums)	conttypenew1
Sickness or Accident	Sickness claim or Accident related claim. (Indicator is for sickness)	sick
Deferred Period	Classified according to defpd0 (0 day), defpd1 (base level and deferment period between 1 and 27 days), defpd2 (28 to 89 day deferment period) and defpd3 (deferment period in excess of 90 days)	defpd0 defpd2 defpd3
Smoker	Indicator variable of smoker; smoker=1	smokernew

We develop a classification scale for insured lives, where 1 indicates extremely short claim duration class and 10 indicates a very long claim duration class. Being able to identify groups of policyholders with similar risks can help actuaries to better understand the risk portfolios underwritten. We attempt to apply and evaluate several data mining methods to predict claim duration class using various rating factors. Results from fitting different prediction models are compared based on two different loss functions. We then suggest using principal component analysis to better understand the relationship between various rating factors and how they impact claim duration class. We also discuss and compare how different variable combining techniques can be used to select and weight available variables used in the prediction models. While we have some understanding of the level of claim termination rates from existing industry tables and other industry level studies (e.g. CMI 12 1991 and Ling et al. 2010), this paper will provide a new outlook on IPI data classifications formulated using modern statistical methods that have not previously been used to advance our understanding of this area in Australia. We hope that this paper will appeal not just to researchers but also to actuarial practitioners in the insurance industry. All models are applied using SAS functions and more details can be found in Liu (2012).

The rest of the paper is organised as follows. Section 2 presents a brief discussion of the data mining models that are used to predict claim duration classes of insured lives. Section 3 demonstrates how principal component analysis and different variable combining techniques can be used to better understand rating factors that drive claim durations. Section 4 presents and compares the results of fitting different data mining models with various variable selection and weighting techniques, and Section 5 concludes the paper.

2 Data Mining Prediction Models

We aim to provide a fast, objective method of scoring claims into different portfolios with homogeneous risks. First of all, we define claim duration as the time difference between claim termination and the date of claim onset. We classify the claim durations T into 10 different classes of duration with approximately 10% of data in each class, ranging from the shortest claim duration class $G = 1$ to the longest claim duration class $G = 10$. The goal is to use various policyholders' information known to the insurers to group the policies into 10 different duration portfolios. The prediction model assigns each new claim a score from 1 to 10- the higher the score, the longer the predicted claim duration.

We apply six different data mining techniques to create our prediction model namely Linear Regression of an Indicator Matrix, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), K-Nearest Neighbour (KNN), Log-logistic Regression and Ordered Log-logistic Regression. Even though we explain some important mathematical details, we emphasise the methods and their conceptual underpinnings more. For more details on model specifications and associated mathematical details see Hastie et al. (2008).

2.1 Linear regression of an indicator matrix

We start off with this most straight forward and widely used regression model. For our analysis, the response variable duration class is denoted by G, which can take values $1, 2, \ldots, 10$. This quantitative outcome is known as a categorical response. By using linear regression, we assume there is no explicit ordering in the classes. We record the predictors using a vector X representing the p predictor variables which include age, gender, and occupation category etc. Components of X are denoted X_j, $j = 1, 2, \ldots, p$. Using N to denote the number of observations, the entire set of the predictors is a set of N input p-vectors $X_i, i = 1, \ldots, N$ recorded in an $N \times p$ matrix $\mathbf{X}$. An indicator variable is used to code each of the response categories. Since G has 10 classes, there will be 10 such indicators $Y_k, k = 1, ..., 10$ with $Y_k = 1$ if $G = k$ else 0. The N training instances of these form an $N \times 10$ indicator response matrix $\mathbf{Y}$. Often it is convenient to include the constant variable 1 in $\mathbf{X}$, and write the linear model as

$$\hat{\mathbf{Y}} = \mathbf{X}^\top \hat{\mathbf{B}},$$

where $\hat{\mathbf{B}}$ denotes the estimated coefficient matrix. There are many different methods to fit the linear model to a set of training data, but by far the most popular is the method of least squares, which minimises the sum of squared residuals and leads to a closed form expression for the estimated parameter $\mathbf{B}$. We fit a linear regression model to each of the columns of $\mathbf{Y}$ simultaneously, and the fit is given by

$$\hat{\mathbf{B}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}.$$

A new observation with input $x^\top = (x_1, \ldots, x_p)$ is classified as follows:

- compute the fitted output $\hat{f}(x)^\top = (1, x^\top)\hat{B}$, a 10 vector;
- identify the largest component and classify accordingly:

$$\hat{G}(x) = \text{argmax}_{k \in G} \hat{f}_k(x),$$

where $\hat{f}_k(x)$ is the fitted linear model for the kth indicator response variable.

Hastie et al. (2008) point out there is a serious masking problem with the regression approach when the number of classes $K \geq 3$, especially prevalent when K is large. Because of the rigid nature of the regression model, classes can be masked by others. We therefore look at Linear Discriminant Analysis (LDA) in the next section, which can prevent this masking problem.

2.2 Linear discriminant analysis

Discriminant analysis is a classic method of classification that has stood the test of time. It was originally developed in 1936 by Fisher. Discriminant analysis often produces models whose accuracy approaches and occasionally exceeds more complex modern methods. Discriminant analysis can only be used for classification problems, not for regression. That is, the target variable must be categorical, but may have two or more categories. Suppose that we model the class density for each observed predictor set using a multivariate Gaussian distribution,

$$f_k(x) = \frac{1}{2\pi^{p/2}|\mathbf{\Sigma}_k|^{1/2}} \exp^{-1/2(x-\mu_k)^\top \mathbf{\Sigma}_k^{-1}(x-\mu_k)}, \text{ for } k \text{ in } 1, 2, \ldots, 10,$$

where p is the number of predictors available, x is the predicting vector with each element representing information such as age, gender, and occupation category etc, μ and $\mathbf{\Sigma}_k$ are the mean vector and variance matrices of the multivariate Gaussian distribution. We are aware of the fact that we have some predictors here that are categorical, which technically violates the normality assumption. However, Manly (1986) pointed out that violation of the normality assumption does not render discriminant analysis a waste of time. He states that discriminant analysis may well turn out excellent even on data from non-normal distributions. We therefore decide to look at both linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA). We will demonstrate in the results section that even though these two models do not perform as well as methods such as k-nearest neighbour, there is still some predictability evident when we look at the misclassification plots.

Let us denote π_k as the prior probability for class k, with $\sum_{k=1}^{10} \pi_k = 1$. A simple application of Bayes' theorem gives us the conditional probability of being in class k conditioning on a set of observed predictors $x^\top = (x_1, x_2, \ldots, x_p)$,

$$\Pr(G = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{10} f_l(x)\pi_l}.$$

LDA arises in the special case when we assume that the classes have the same covariance matrix $\mathbf{\Sigma}$ for each class k. The method is called linear discriminant analysis because the decision boundaries between any two classes are linear in the predictor vector x. We can see this through the log ratio of the decision boundary for classes k and l,

$$\begin{aligned}\log\frac{\Pr(G=k|X=x)}{\Pr(G=l|X=x)} &= \log\frac{f_k(x)}{f_l(x)}+\log\frac{\pi_k}{\pi_l} && (1)\\ &= \log\frac{\pi_k}{\pi_l}-\frac{1}{2}(\mu_k+\mu_l)^\top\mathbf{\Sigma}^{-1}(\mu_k-\mu_l)+x^\top\mathbf{\Sigma}^{-1}(\mu_k-\mu_l). && (2)\end{aligned}$$

The predicted outcome is classified based on the decision function

$$G(x)=\text{argmax}_k\delta_{k(x)},$$

where $\delta_{k(x)}$ are the conditional probabilities of being classified as class k given a set of predictors x:

$$\delta_k(x)=\log\{\Pr(G=k|X=x)\}=x^\top\mathbf{\Sigma}^{-1}\mu_k-\frac{1}{2}\mu_k^\top\mathbf{\Sigma}^{-1}\mu_k+\log\pi_k. \quad (3)$$

These linear discriminant functions are equivalent to the decision rule described in Equation (2). In practice, the parameters μ and $\mathbf{\Sigma}$ in the Gaussian distributions and the prior probability π_k are not known and therefore we need to estimate them using our training data. The algorithm is as follows:

- Estimate the prior probability for each class $\hat{\pi}=N_k/N$, where N_k is the number of observations in class k and N is the total observations in the training set.
- Estimate the mean for each class: $\hat{\mu}_k=\Sigma_{y_i=k}x_i/N_k$, $k=1,2,\dots,10$.
- Estimate the common covariance matrix $\hat{\mathbf{\Sigma}}=\Sigma_{k=1}^{10}\Sigma_{y_i=k}(x_i-\hat{\mu}_k)(x_i-\hat{\mu}_k)^\top/(N-10)$.

With two classes there is a simple correspondence between linear discriminant analysis and classification by linear regression introduced in earlier section. Hastie et al. (1994) showed that with more than two classes, LDA is not the same as linear regression of the class indicator matrix, and it avoids the masking problems associated with that approach.

2.3 Quadratic discriminant analysis

Quadratic Discriminant Analysis (QDA) is similar to LDA, except that separate covariance matrices $\mathbf{\Sigma}_k$ must be estimated for each class of outcomes. If $\mathbf{\Sigma}_k$ are not assumed to be the same for all 10 classes, the quadratic part in the exponents of Equation (2) do not cancel out, and the decision boundary is therefore a quadratic function of the predictors x. Applying similar logic to Equation (3), we get the quadratic discriminant functions,

$$\delta_k(x) = -\frac{1}{2}\log|\mathbf{\Sigma}_k| - \frac{1}{2}(x-\mu_k)^\top \mathbf{\Sigma}_k^{-1}(x-\mu_k) + \log\pi_k \text{ for } k = 1, 2, \ldots, 10.$$

The decision boundary between each pair of classes k and l is described by a quadratic equation $\{x : \delta_k(x) = \delta_l(x)\}$.

Michie et al. (1994) demonstrated that both LDA and QDA perform well on a large and diverse set of classification tasks. Hastie et al. (2008) suggest that the reason why LDA and QDA to have such good performance records is not because the data are approximately Gaussian, or the covariances are approximately equal for LDA. They suggest it is because the data can only support simple decision boundaries such as linear or quadratic, and the estimates provided via the Gaussian models are stable. That is, these models may have relatively high bias, but can be estimated with much lower variance than more complicated alternatives.

2.4 Multinomial logistic regression

Logistic regression is another type of regression analysis used for predicting the categorical outcomes based on one or more predictor variables. For categorical outcome it is inappropriate to use linear regression because the linear regression model can generate any real number ranging from negative infinity to positive infinity, whereas our categorical outcome can only take on discrete values $1, 2, \ldots, 10$. Instead of equating the expected value of the dependent variable to a linear combination of independent variables and their corresponding parameters like linear regressions do, logistic regression models equate the linear component to the logit function of the probability of a given outcome on the dependent variable. For our analysis, we have 10 discrete duration classes for the dependent variable. We will consider the 10th category to be the omitted or baseline category, where logits of the first 9 categories are constructed with the baseline category in the denominator. The model has the form

$$\log\frac{\Pr(G = k|X = x)}{\Pr(G = 10|X = x)} \quad = \quad \alpha_k + \beta_k^\top x, \text{ for } k \text{ in } 1, 2, \ldots, 9,$$

where x still denotes the observed predicting vector $x^\top = (x_1, x_2, \ldots, x_p)$, and α_k and β_k are the intercepts and regression coefficient vector respectively. Although we use class 10 as the denominator in the odds-ratios, the choice of denominator is arbitrary. Solving for $\Pr(G =$

$k|X = x)$, we have

$$\Pr(G = k|X = x) = \frac{\exp(\alpha_k + \beta_k^\top x)}{1 + \sum_{l=1}^{9} \exp(\alpha_l + \beta_l^\top x)}, \text{ for } k = 1, 2, \ldots, 9$$

$$\Pr(G = 10|X = x) = \frac{1}{1 + \sum_{l=1}^{9} \exp(\alpha_l + \beta_l^\top x)}.$$

The probabilities of being classified in each of the 10 classes obviously sum to one. Denote G to be the duration class random variable, which can take one of 10 possible values. For a given set of predictors x on N observations, G can be considered a column vector of N multinomial random variables G_i. Let $\mathbf{Y}$ be a $N \times 10$ indicator response matrix with $Y_{ik} = 1$ if $G_i = k$ else 0. $\boldsymbol{\pi}$ is a matrix of the same dimension as $\mathbf{Y}$ where each element π_{ik} is the probability of observing the kth value of the dependent variable G for any given observation in the ith row. Thus, the likelihood function is

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}|\mathbf{y}) = \prod_{i=1}^{N} \prod_{k=1}^{10} \pi_{ik}^{y_{ik}}.$$

The log-likelihood is

$$l(\boldsymbol{\alpha}, \boldsymbol{\beta}|\mathbf{y}) = \sum_{i=1}^{N} \sum_{k=1}^{10} \log \pi_{ik}^{y_{ik}}.$$

Unlike linear regression with normally distributed residuals, there is no closed-form expression for the coefficient values that maximise the likelihood function, however we can use an iterative process such as Newton's method instead.

By fitting a multinomial logistic model, we ignore any ordering in the values of the duration class. We fit the same model if the duration class is as short as 1 or as long as 10. The advantage of the logistic regression model is that we estimate coefficients that capture differences between all possible pairs of duration classes. However, for our analysis, the predicted outcome is clearly ordered from short duration to long duration, therefore we also look at ordered logistic regression in the next section.

2.5 Ordered logistic regression model

Our dependent variable duration class is ordinal, that is, it is ranked from the shortest duration class to the longest duration class. We consider a model that incorporates the ordinal nature of our dependent variable. The ordered logistic regression model, also known as the proportional odds model can be applied here. Ordered logistic regression models cumulative probability. For our analysis, the event of interest is observing a particular duration class or less. We model the following odds ratios (McCullagh, 1980) :

$$\log \frac{\Pr(G \leq k|X = x)}{1 - \Pr(G \leq k|X = x)} = \alpha_k - \beta^\top x, \text{ for } k = 1, \ldots, 9,$$

where all the parameters and vectors are defined in the same way as logistic model defined in Section 2.4. The last category does not have an odds associated with it since the probability of being classified into a duration group that is less than or equal to 10 is 1. Notice the negative sign of the β coefficients in the linear predictors above. This ensures that larger coefficients indicate an association with a longer duration group. For a continuous variable, a positive coefficient tells us that as the values of the variable increase, the likelihood of being in a longer duration group increases. An association with longer duration class means smaller cumulative probabilities for shorter duration class, since they are less likely to occur. The model is called the proportional odds model because the log of cumulative odds ratio of making the same responses at different x-points is proportional to the distance of the points,

$$\log \left\{ \frac{\Pr(G \leq k|X = x_1)}{1 - \Pr(G \leq k|X = x_1)} \times \frac{1 - \Pr(G \leq k|X = x_2)}{\Pr(G \leq k|X = x_2)} \right\} = \beta^\top (x_2 - x_1). \tag{4}$$

Maximum likelihood estimation is used to estimate the parameters α_k and β. Using the relationship equation $\Pr(G = k|X = x) = \Pr(G \leq k|X = x) - \Pr(G \leq (k-1)|X = x)$, the likelihood function can be derived in a similar way as for the logistic regression model.

The model constrains the classified group curves to have the same shape as shown in Equation (4), and therefore we cannot fit it by fitting separate logit models for each group as for multinomial logistic regression model introduced in 2.4. We must maximise the multinomial likelihood subject to a constraint. The model only applies to data that meet the proportional odds assumption, that is the relationship between any two pairs of outcome groups is statistically the same.

2.6 Nearest-neighbour classifier

The k-Nearest-Neighbour Method is essentially a model-free method for classification and pattern recognition. Because it is highly unstructured, it typically is not useful for understanding the nature of the relationship between the predictors and class outcome. However, as a black box prediction engines, it can be very effective, and is often among the best performers in real data problems. The k-nearest classifiers are memory-based, given a query point x_0, we find the k training points $x_{(r)}, r = 1, \ldots, k$ closest in distance to x_0, and then classify using majority vote among the k neighbors. Ties are broken at random. For simplicity we will assume that the features are real-valued, and we use Euclidean distance in the feature space:

$$d_i = |x_i - x_0|.$$

The k-nearest neighbour fit for duration class $\hat{Y}$ is defined as follows:

$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i,$$

where $N_k(x)$ is the neighbourhood of x defined by the k closest points x_i in the training sample, and we average the their corresponding observed class indicator outcomes. We then find the largest fitted value among $\hat{Y}_1, \ldots, \hat{Y}_{10}$, and assign to class $\hat{G}$ accordingly. It is not hard to understand that for k-nearest-neighbour fits, the error on the training data should be approximately an increasing function of k. The bias of the 1-nearest-neighbor estimate is often low because it uses only the training point closest to the query point, but the variance is high. A famous result of Cover and Hart (1967) shows that asymptotically the error rate of the 1-nearest-neighbor classifier is never more than twice the Bayes rate. We would need an independent test set for comparing the different methods.

Unlike the least squares methods, the k-nearest-neighbour procedures do not appear to rely on any stringent assumptions about the underlying data. However, any particular subregion positions of the decision boundary depends on a handful of input points and is thus too specific to the data and unstable, resulting high variance and low bias. It seems that with a reasonably large set of training data, we could always approximate the theoretically optimal conditional expectation by k-nearest-neighbour averaging, since we should be able to find a fairly large neighbourhood of observations close to any x and average them. However this approach breaks down in high dimensions. It can be shown that, in a high dimensional space, in order to capture a small percentage of the data to form a local average, we must cover a big range of each input variable. Such neighbourhood is no longer “local”. This is commonly known as the curse of dimensionality (Bellman, 1961).

2.7 Model assessment

We discuss the appropriate performance assessment criteria to use in order to compare models. A loss function approach is often used to evaluate the accuracy of a categorical predictor $\hat{G}$. Let $L(G;\hat{G})$ denote the loss incurred when $\hat{G}$ is used to predict a random variable G . A loss function $L(.)$ usually satisfies the following conditions: it is bounded below by 0 and attains 0 when correct prediction is made, i.e. $\hat{G} = G$; For a categorical variable G, a widely used loss function is:

$$L(G;\hat{G}) = I(G \neq \hat{G}) \tag{5}$$

where $I(.)$ is the indicator function. For a set of test data, the prediction error e can be estimated via

$$\hat{e} = L_{\text{test}} = \frac{1}{M}\sum_{i=1}^{M} L_i$$

where M is the number of test data, and L_i is a indicator variable for misclassification. One disadvantage of this loss function is that it does not measure how far off the misclassification is. In other words, misclassifying a $G = 1$ observation as $G = 10$ has the same contribution to prediction error as misclassifying it to $G = 2$. We prefer the loss function to increase as the distance between G and $\hat{G}$ increases. A more appropriate loss function for our analysis is therefore

$$L(G;\hat{G}) = |G - \hat{G}|, \tag{6}$$

and L is the loss distance between the number of class difference between the test observation and the predicted class outcome.

3 Variable Selection

We mentioned in earlier sections that some data mining models such as k-nearest neighbour does not provide much useful information on the relationship between the predictors and the class outcome. We can use methods such as forward selection or backward elimination to select variables (see for example, Hocking, 1976; Wilkinson and Dallal, 1981). Forward selection involves starting with no variables in the model, testing the addition of each variable using a chosen model comparison criterion, adding the variable that improves the model the most, and repeating this process until none improves the model. Backward elimination involves starting with all candidate variables, testing the deletion of each variable using a chosen model comparison criterion, deleting the variable that improves the model the most by being deleted, and repeating this process until no further improvement is possible. However the processes of these methods can be time consuming especially when there are many variables available. We therefore propose two ways of selecting and weighting available variables. We suggest that both principal component analysis and weighted combination models can be used to understand the role of the predicting variables in explaining claim duration class, and possibly detect variables with significant interaction.

3.1 Principal Component Analysis Model

The technique of principal component analysis was first described by Karl Pearson (1901), although he did not propose a practical method of calculation for more than two or three variables. Hotelling (1933) described the computing methods. The object of the analysis is to take p variables $X_1, X_2, \ldots, X_p$ and find combinations of these to produce components $Z_1, Z_2, \ldots, Z_p$ that are uncorrelated. Since $Z_1, Z_2, \ldots, Z_p$ are uncorrelated, it means they measure different dimensions in the data. The components are also ordered so that Z_1 displays the largest amount of variation in the response, Z_2 displays the second largest amount of variation, and so on. We aim to explain the variation in the data set using as few Z variables with variances that are not negligible. If a smaller number of Z variables can account for the variation in the p original X variables, we can achieve some degree of economy without worrying about which of the original X variables are correlated or significant for the model under consideration. The method is also useful if we think there is a good deal of redundancy in the original variables, with most of them measuring similar things. The principal component algorithm is as follows:

- Standardise $X_1, \ldots, X_n$ to have zero means and unit variances.

- Compute the correlation matrix $\mathbf{C} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^\top$ for the original variables. This is a correlation matrix because X_i are standardised already.

- Compute the variances of the principal components are the eigenvalues of the matrix C. Assuming that the eigenvalues are ordered as $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$, then λ_i correspondes to the ith principal component. Find eigenvalues: $\lambda_1, \lambda_2, \ldots, \lambda_p$ and the corresponding eigenvectors $a_1, a_2, \ldots, a_p$. The coefficients of the ith principal component are then given by a_i while λ_i is its variance.

- Discard any components that only account for a small proportion of the variation in the response.

3.2 Principal Component Analysis Results

There are a total of 16 variables available that have possible impact on claim duration. It can be argued as to how many components to choose is appropriate. One rule is to use the mineigen criterion, which states that only components with eigenvalues above 1 should be retained. Components with an eigenvalue of less than 1 account for less variance than did the original standardised variable, and so are of little use. In Table 2, the first 6 principal components have eigenvalues over 1. We can use these six principal components for claim duration class prediction later on. Due to the dropping of the less important components, the sum of these six components can only explain a total variance of 8.71. This is also called the final communality estimate, which is off course less than total variance 16 of the original correlation matrix. Notice that the six most significant components only account for around 54% of the total variance in the original correlation matrix. We will later on see in the Results section that if we only use these six most significant extracted components as predictors in the data mining model, we will not get as good predictability as using all available variables due to the information loss.

Table 2: Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	1.9949	0.3113	0.1247	0.1247
2	1.6836	0.1763	0.1052	0.2299
3	1.5073	0.2092	0.0942	0.3241
4	1.2980	0.1536	0.0811	0.4052
5	1.1444	0.0621	0.0715	0.4768
6	1.0823	0.1039	0.0676	0.5444
7	0.9784	0.0407	0.0611	0.6055
8	0.9376	0.0266	0.0586	0.6642
9	0.9110	0.0225	0.0569	0.7211
10	0.8885	0.0703	0.0555	0.7766
11	0.8182	0.0545	0.0511	0.8278
12	0.7637	0.0887	0.0477	0.8755
13	0.6750	0.0936	0.0422	0.9177
14	0.5814	0.0630	0.0363	0.9540
15	0.5184	0.3010	0.0324	0.9864
16	0.2174		0.0136	1.0000

We can normally get some useful insights on the rating factors by looking at the factor pattern output, which is often referred to as the factor loading matrix in principal component analysis. The elements in the loading matrix are called factor loadings. There are at least two ways we can interpret these factor loadings. First, we can use this table to express the observed variables as functions of the extracted components. Each row of the factor loadings tells us the linear combination of the component scores that would yield the expected value of the associated variable. Second, we can interpret each loading as a correlation between an observed variable and a component, provided that the factor solution is an orthogonal one, that is, components are uncorrelated, such as the current initial factor solution. Hence, the factor loadings indicate how strongly the variables and the components are related. Values greater than 0.30122 are flagged by an *, which means the corresponding components have the largest loadings on these variables. In Table 3, the first component labelled component 1 is highly correlated with $poldesnew3$ and $benrate$, but in opposite directions. $Poldesnew3$ is an indicator variable for any occupation disability definition. Any occupation means an income payout occurs if by reason of illness, accident or injury the insured is unable to perform any work at all. Own Occupation definition means an income payout occurs if by reason of illness, accident or injury the insured live is unable to perform his/her normal own occupation that the insured person is reasonably suited by education, training or experience. The data show that higher benefit rate and own occupation tend to have a large positive impact on component 1. We can therefore interpret component 1 as a score for the degree of professionalism. A high degree of professionalism tends to be associated with high salary and therefore a higher benefit rate. It is intuitive that people whose occupation are highly professional will generally take out cover for own occupation rather than any occupation. It is quite obvious that component 2 identifies occupation class C rather than occupation class D. Component 6 identifies claims with long deferred period (defpd3 is 1) that tend to be censored. Similarly, component 4 captures claims due to sickness, which tend to have a short deferred period (deferred period being 0).

Table 3: Rotated Factor Pattern

	Component 1	Component 2	Component 3	Component 4	Component 5	Component 6
age	-1	-2	65 *	-31 *	-1	14
terminate	0	-3	3	7	-12	-64 *
poldesnew3	-72 *	-1	-3	10	-10	8
sex1	-9	-6	9	24	-66 *	3
occupB	-1	-8	15	25	72 *	7
occupC	-19	87 *	-12	4	-26	-8
occupD	-28	-82 *	-16	-1	-30	-8
benrate	70 *	-1	19	9	-16	18
bentypnew2	53 *	14	-43 *	11	6	13
medevid1	-15	-11	41 *	4	15	11
conttypenew1	7	22	54 *	6	-2	8
smokernew	-14	6	-53 *	-3	5	31 *
sick	1	3	17	-71 *	18	17
defpd0	-7	7	4	78 *	17	6
defpd2	52 *	-4	-13	-33 *	20	-10
defpd3	6	-4	13	2	-10	70 *

In Table 4, each component is expressed as a linear combination of the standardised observed variables. For example, the first principal component is computed as:

$$\begin{aligned}
&-0.0158 \ast \text{age} + 0.0481 \ast \text{terminate} - 0.4239 \ast \text{poldesnew3} + 0.0364 \ast \text{sex1} \\
&-0.0490 \ast \text{occupB} - 0.1373 \ast \text{occupC} - 0.0963 \ast \text{occupD} + 0.4459 \ast \text{benrate} \\
&+0.2993 \ast \text{bentypnew2} - 0.0868 \ast \text{medevid1} + 0.0464 \ast \text{conttypenew1} \\
&-0.1278 \ast \text{smokernew} - 0.0763 \ast \text{sick} - 0.0011 \ast \text{defpd0}.
\end{aligned}$$

Notice that, when applying this formula we must use the standardised observed variables (with means 0 and standard deviations 1), but not the raw data. We can then apply each data mining methods using these component scores as predictors. Results are provided and compared with other variable selection methods later on in the next section.

Table 4: Standardised Scoring Coefficients

	Component 1	Component 2	Component 3	Component 4	Component 5	Component 6
age	-0.0158	0.0023	0.4054	-0.1700	-0.0469	0.0731
terminate	0.0481	-0.0139	0.0799	0.0430	-0.0590	-0.5617
poldesnew3	-0.4239	0.0275	-0.0374	0.0049	-0.0055	0.1123
sex1	0.0364	-0.0484	0.0991	0.1577	-0.5163	0.0625
occupB	-0.0490	-0.0643	0.0810	0.2013	0.5696	0.0160
occupC	-0.1373	0.5858	-0.0678	-0.0383	-0.1763	-0.0478
occupD	-0.0963	-0.5229	-0.0997	-0.0117	-0.1997	-0.0186
benrate	0.4459	-0.0541	0.1449	0.1348	-0.2173	0.1244
bentypnew2	0.2993	0.0552	-0.2755	0.0906	0.0048	0.1211
medevid1	-0.0868	-0.0649	0.2584	0.0537	0.1085	0.0642
conttypenew1	0.0464	0.1380	0.3650	0.0766	-0.0460	0.0242
smokernew	-0.1278	0.0420	-0.3909	-0.0696	0.0607	0.3210
sick	-0.0763	0.0493	0.0478	-0.4796	0.1145	0.1124
defpd0	-0.0011	0.0176	0.0699	0.5454	0.1515	0.0600

3.3 Combination Prediction

Combination forecast is a technique that has been adopted by economics and finance researchers to forecast future stock returns. As pointed out in the seminal paper by Bates and Granger (1969), combinations of individual forecasts can outperform the individual forecasts themselves. Forecast combination has recently received renewed attention in the macroeconomic forecasting literature with respect to forecasting inflation and real output growth (e.g., Stock and Watson 1999, 2003, 2004). We explore the combinations of individual forecast idea and apply it to the out-of-sample class prediction. Suppose there are N observations in the training data with p predicting variables available. First of all, for each observation i in the training data, we apply the data mining technique using each of the individual variables to get p estimated class predictors $\hat{G}_j$ for $j = 1, \ldots, p$. Then for variable j, we calculate estimated weight W_j as the reciprocal of the loss value L_j defined by Equation (6),

$$W_j = \frac{1}{L_j} \text{ for } j \text{ in } 1, \ldots, p. \tag{7}$$

The higher the loss value, the smaller weight we wish to place on that variable. The next step is to calculate the average weight $\bar{W}_j$ for variable j based on the N observations in the training data,

$$\bar{W}_j = \frac{1}{N} \sum_{i=1}^{N} W_j;$$

The estimated combination predictor $\hat{G}$ is taken to be the weighted average of the p individual estimated predictor based on the following equation:

$$\hat{G} = \sum_{j=1}^{p} \bar{W_j} * \hat{G}_j$$

where $\hat{G_j}$ is the predicted class using one of the data mining techniques with an individual variable j, and $\bar{W_j}$ is the ex ante combining weight which is the average weight used for that particular individual prediction for the training observations. We are going to refer this method as weighted combination prediction method in later sections. Table 5 provides $\bar{W_j}$ results for j in $1, \ldots, p$ based on different data mining methods. We flag the six variables assigned the largest weights by $*$ to indicate the variables that are found most significant under different data mining models. Own occupation definition ($poldesnew3$), Occupation class D indicator ($occupD$), and deferred period 0 indicator ($defpd0$) are among the six most significant variables across all models considered. If we compare this finding with the results found by principal component analysis in Table 3, it is interesting to see that the same variables were found to be highly correlated to the most significant components. Therefore the two methods provide fairly consist results in terms of identifying variables that are significant.

Table 5: Weights Table

	age	terminate	poldesnew3	sex1	occupB	occupC	occupD	
KNN	0.0650	0.0589	0.0654	0.0364	0.0956	0.0569	0.0805	
LDA	0.0498	0.0650	0.0672	0.0402	0.1088	0.0555	0.0808	
Linear	0.0558	0.0592	0.0705	0.0642	0.0585	0.0525	0.0756	
Loglog	0.0596	0.0633	0.0744	0.0684	0.0624	0.0558	0.0796	
OLL	0.0672	0.0661	0.0686	0.0634	0.0628	0.0690	0.0697	
QDA	0.0553	0.0635	0.0674	0.0388	0.1041	0.0586	0.0814	
	benrate	bentypnew2	medevid1	conttypenew1	smokernew	sick	defpd0	defpd2
KNN	0.0663	0.0723	0.1047	0.0592	0.0346	0.0585	0.1029	0.0429
LDA	0.0464	0.0722	0.1053	0.0663	0.0387	0.0576	0.1035	0.0427
Linear	0.0564	0.0554	0.0574	0.0561	0.0570	0.0522	0.1147	0.0565
Loglog	0.0598	0.0586	0.0613	0.0598	0.0606	0.0553	0.1210	0.0601
OLL	0.0642	0.0683	0.0622	0.0632	0.0640	0.0691	0.0724	0.0697
QDA	0.0415	0.0731	0.1064	0.0633	0.0371	0.0602	0.1044	0.0448

Rapach et al. (2010a) also explained two other more simple combining methods for combination forecast. The methods only differ in how the weights are determined. One of them uses simple averaging mean weight. The mean combination prediction sets $W_j = 1/p$, for $j = 1, \ldots, p$. Therefore each individual variable predictor gets the same weight when combined together. The other method is called the trimmed mean combination forecast, which sets

$W_j = 0$ for the individual prediction with the smallest and largest value and $W_j = 1/(p-2)$ for the remaining individual prediction. We will apply the data mining models using all of the above mentioned variable combining methods to predict duration classes and compare their performances in the next section.

4 Data Mining Prediction Model Results

In order to build and test the model, we randomly divide the data into training and test sets according to the ratio of 80% to 20%. Data mining models are applied to the training data and a prediction model is constructed. As the test data are not used in the training of a model, they provide an independent way to evaluate the model. The test data are also used to compare and rank various models considered. For each data mining method, we obtain five sets of results. Firstly, we apply the data mining methods with all 16 available variables to do out-of-sample duration class prediction. Secondly, we apply the principal component analysis to extract a few significant components that explain most variation in the variables, we then apply the data mining techniques using the principal components found. The last three approaches we adopt are the weighted combination prediction, the mean combination prediction, and the trimmed mean combination prediction. We apply the data mining method using each individual variable and then use these three different combining methods to predict claim class for test data.

Table 6 to Table 8 provide the results on the prediction power of each data mining method. Total loss is defined as the total number of mis-classifications based on the test data using the modified loss function defined by Equation (6). The average number of mis-classifications together with the standard error of the mis-classifications are also given. Notice that, these are all based on the modified loss function defined by Equation 6. If we define the loss function using Equation (5), it would result in very high prediction error over 80%, which overlooks the true prediction power of data mining. The predictability of data mining methods is clearly evident in the plot for misclassification. In Figure 1 to Figure 6, it is demonstrated that for every data mining method with every variable selection approach, there is a clear decreasing trend of policy numbers as the number of misclassifications increases. This suggests that even though the data mining methods cannot get the duration class prediction exactly right, most of the predicted outcomes are close to what we expect. This predictability can only be detected if we use the loss function defined by Equation (6).

Table 6: Data Mining Results with All variables

	total loss	mean loss	standard deviation	prediction error
linear regression	5028	2.8375	2.4556	0.8053
linear discriminant analysis	4365	2.4633	2.1555	0.8138
quadratic discriminant analysis	5241	2.9577	2.4605	0.8200
k-nearest neighbour	4434	2.5023	2.1889	0.8121
loglogistic regression	5003	2.8234	2.4351	0.8070
ordered loglogistic regression	5112	2.8849	2.4187	0.8070

Table 7: Data Mining Results with principal Components

	total loss	mean loss	standard deviation	prediction error
linear regression	5432	3.0655	2.5336	0.8222
linear discriminant analysis	4882	2.7553	2.3271	0.8149
quadratic discriminant analysis	4739	2.6744	2.2351	0.8442
k-nearest neighbour	4576	2.5824	2.1488	0.8358
loglogistic regression	5055	2.8527	2.3995	0.8188
ordered loglogistic regression	5336	3.0113	2.3755	0.8369

Table 8: Data Mining Results with Weighted Combination Prediction

	total loss	mean loss	standrad deviation	prediction error
linear regression	4621	2.6078	1.8118	0.8888
linear discriminant analysis	4122	2.3262	1.4732	0.8883
quadratic discriminant analysis	4127	2.3290	1.4677	0.8900
k-nearest neighbour	4091	2.3087	1.4598	0.8860
loglogistic regression	4570	2.5790	1.7792	0.8900
ordered loglogistic regression	4377	2.4701	1.6805	0.8860

Table 9: Data Mining Results with Mean Combination Prediction

	total loss	mean loss	standard deviation	prediction error
linear regression	4929	2.7816	1.9653	0.8911
LDA	4122	2.3262	1.4666	0.8916
QDA	4144	2.3386	1.4834	0.8928
KNN	4101	2.3143	1.4567	0.8866
loglogistic regression	4748	2.6795	1.9051	0.8821
Ordered loglogistic regression	4420	1.7051	1.7051	0.8900

Table 10: Data Mining Results with Trimmed Mean Combination Prediction

	total loss	mean loss	standard deviation	prediction error
linear regression	5369	3.0299	2.1551	0.8900
LDA	4260	2.4041	1.5684	0.8939
QDA	4281	2.4159	1.5792	0.8945
KNN	4187	2.3629	1.5311	0.8933
loglogistic regression	5369	3.0299	2.1551	0.8900
Ordered loglogistic regression	4651	2.6247	1.8583	0.8775

Figure 1: Misclassification plot Using Linear Regression Model

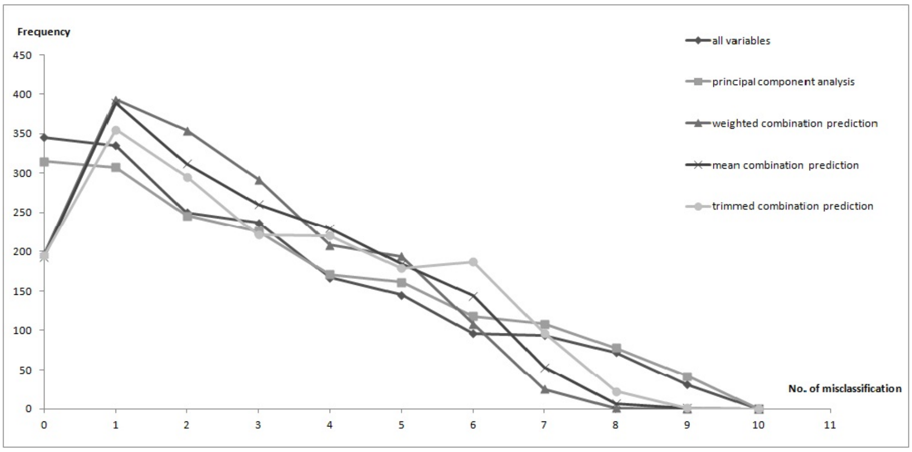


Figure 2: Misclassification plot Using Linear Discriminant Model

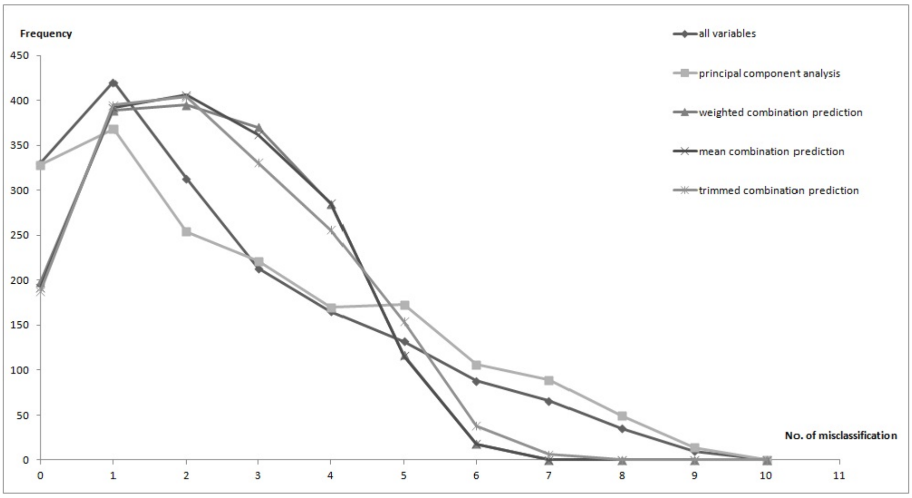


Figure 3: Misclassification plot Using Quadratic Discriminant Model

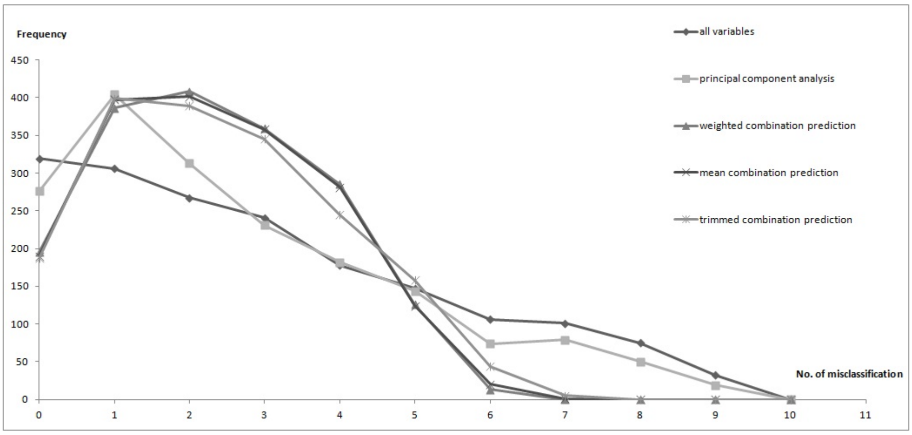


Figure 4: Misclassification plot Using K-nearest Neighbour Model

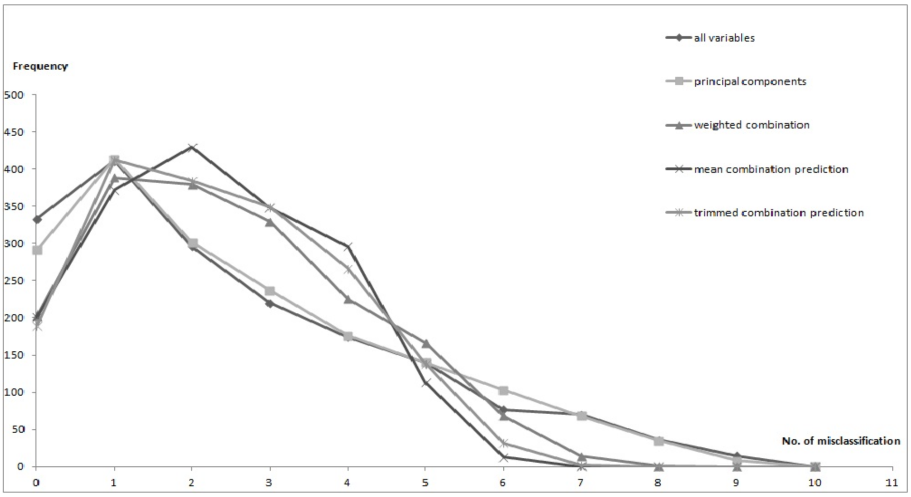


Figure 5: Misclassification plot Using Logistic Model

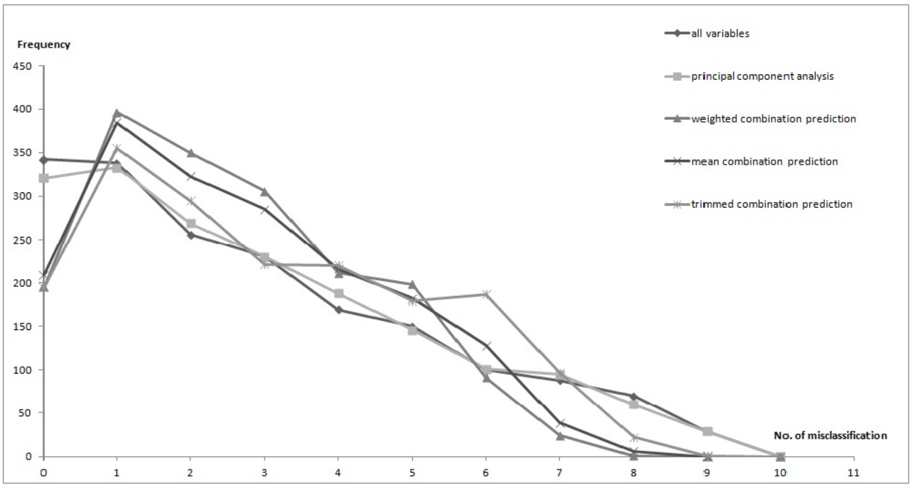


Figure 6: Misclassification plot Using Ordered Logistic Model

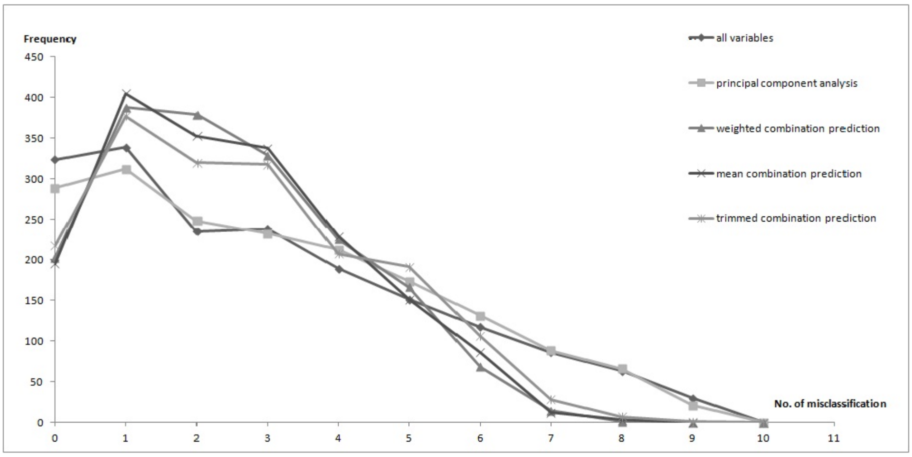


Figure 7: Total Loss on Different Data Mining Models

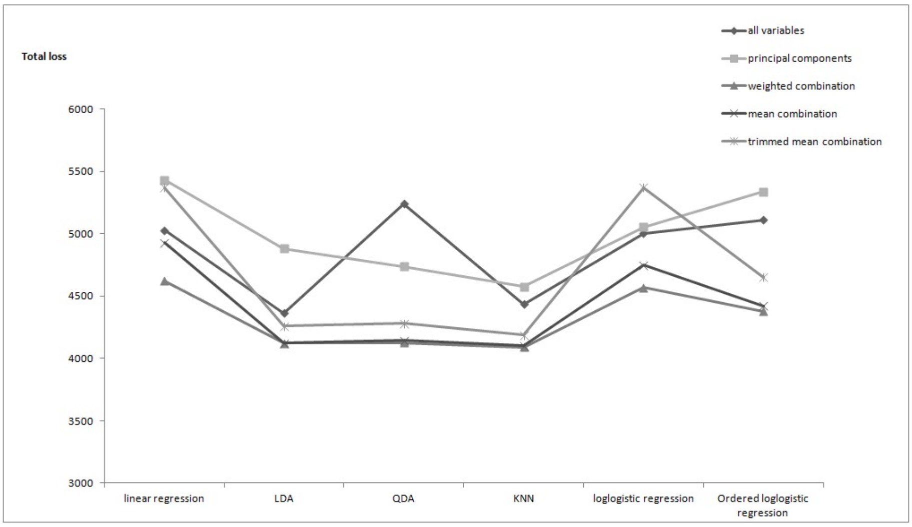


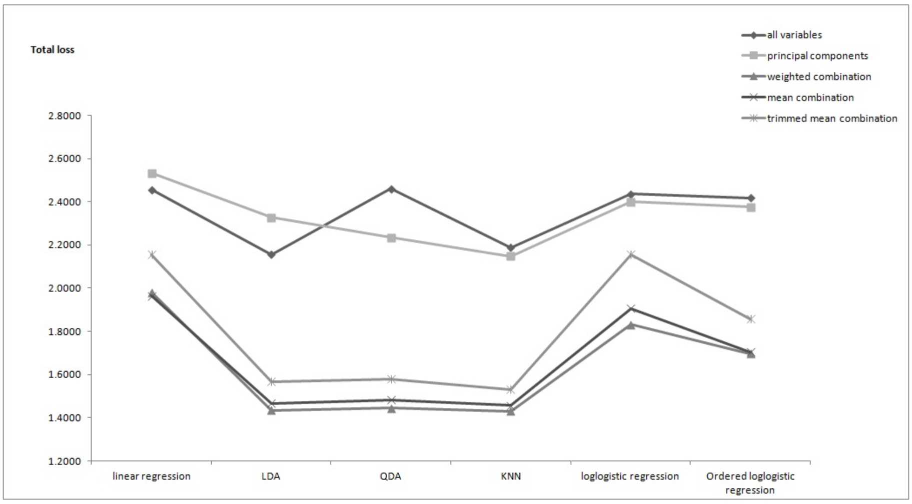
Figure 7 provides the total loss plots for different data mining models together with various variable selection methods. It is a good way to compare across different data mining models or to compare the variable selection methods we have used. There are a few interesting findings based on Figure 7.

- First of all, if we look along each line, k-nearest neighbour always gives the minimum loss value for all six variable selection methods compared to the other data mining models. This proves that as a black box prediction engine, k-nearest neighbour can be very effective, and is often among the best performers in real data problems. However, as we pointed out earlier, because it is highly unstructured, it is hard to understand the nature of the relationship between the predictors and class outcome. To better understand the significance and relationship of various available rating factors, we can use principal component analysis or look at the weighting results from the weighted combination method explained earlier. Secondly, if we compare across different variable selection/weighting methods, the weighted combination method is the lowest line, meaning it consistently outperforms the other five variable selection methods no matter which data mining model we use. This out performance is also clearly evident in the misclassification histogram we saw previously in Figure 1 to Figure 6. The weighted combination prediction performs especially well for eliminating very large misclassification, that is misclassification being 7 to 10 when compared with other variable weighting methods. Again, this strong predictability of weighted combination method is masked when we only look at the conventional prediction error provided in Table 6 to Table 10 calculated based on Loss Equation (5). Even though the weighted combination method did not predict the duration class perfectly, that is having fewer misclassifications being zero compared to the other variable selection methods, it does a very good job in getting the predicted outcome close to the true duration class rather than getting it far off track. Therefore, we stress the point that using the modified loss function from Equation (6) gives more information while assessing data mining performance.

- Finally, if we compare the first approach (all variables) with the second approach (principal component analysis), using only six principal components extracted as predictors for data mining models performs worse than using all variables available most of the time. This is what we expected as we have seen in Table 2, the six most significant components only account for around 54% of the total variance in the original correlation matrix. Nevertheless, it was still useful to look at the factor loading matrix provided in Table 3 to better understand how different rating factors impact duration classes.

We have also plotted the standard deviations of loss values for different data mining models considered. Again, the weighted combination method is the best performer in terms of having the smallest standard deviation. Comparing along each line, k-nearest neighbour is always among the best two data mining methods in terms of having the smallest loss standard deviation across different variable selection methods.

Figure 8: Loss Standard Deviation for Different Data Mining Models



5 Conclusion

This paper demonstrates how data mining can be applied to Income Protection Insurance data to classify insured lives into portfolios with homogeneous risks. We provide a fast, objective method of scoring claims into different claim duration classes. Being able to identify groups of policyholders with similar risks can help actuaries to better understand the risk portfolios underwritten. Results from fitting different prediction models are compared based on two different loss functions. The predictability of all the data mining methods considered is clearly evident when we look at the plot of misclassification. However this predictability can be masked if we only look at the conventional prediction error rate. k-nearest neighbour was found to be the best performer among all the different data mining models considered in terms of having the smallest mean loss value and the smallest loss standard deviation no matter which variable selection or combining method we use. However one of the limitations of such a black box model is that there is not much useful information provided for understanding the relationship between rating factors. We therefore suggest principal component analysis as a way of understanding factor patterns. Moreover, we discuss and compare how different combination models can be used to weight available predicting variables. Principal component analysis and weighted combination prediction model provide very consistent results on identifying the significant variables in explaining claim durations. We find that the occupation definition used to assess whether a policyholder is eligible for a benefit payment, occupation class and deferred period are the most important information to predict claim duration class. All in all, we suggest that data mining techniques can provide some useful insights for informing claim termination rate estimation, and this paper should appeal to both researchers and practitioners.

References

[1] Bates, J.M. and Granger, C.W.J., 1969. The combination of forecasts. Operational Research Quarterly, 20, 451-468.

[2] Bellman, R. E., 1961. Adaptive Control Processes, Princeton University Press.

[3] Cover, T. and Hart, P. 1967 Nearest neighbor pattern classification, IEEE Transactions on Information Theory IT-11: 21-27.

[4] Efron, B., 1975. The efficiency of logistic regression compared to normal discriminant analysis, Journal of the American Statistical Association 70: 892-898.

[5] Fisher, R. A., 1936. The Use of Multiple Measurements in Taxonomic Problems, Annals of Eugenics 7 (2): 179C188.

[6] Hocking, R. R., 1976. The Analysis and Selection of Variables in Linear Regression, Biometrics, 32.

[7] Hotelling, H., 1933 Analysis of a complex of statistical variables into principal components, Journal of Educational Psychology, 24, 417-41 and 498-520.

[8] Hastie T., Tibshirani R., and Friedman J., 2008. The Elements of Statistical Learning: Data Mining, Inference, and Prediction

[9] Ling S.Y., Waters H.R., and Wilkie A.D., 2010. Modelling Income Protection Insurance claim termination rates by cause of sickness I: Recoveries, The Annals of Actuarial Science, 4 (2): 199 - 240.

[10] Liu Q., Statistical modelling of insurance claims, Doctor of Philosophy Thesis, Faculty of Business and Economics, The University of Melbourne.

[11] Manly, 1986. Multivariate Statistical Methods: A Primer, Chapman Hall, London.

[12] McCullagh, P., 1980 Regression Models for Oridinal Data, Journal of the Royal Statistical Society, Series B (Methodological), 42(2), 109-142.

[13] Michie, D., Spiegelhalter, D. and Taylor, C. (eds), (1994). Machine Learning, Neural and Statistical Classification, Ellis Horwood Series in Artificial Intelligence, Ellis Horwood.

[14] Pearson, K., 1901 On lines and planes of closest fit to a system of points in space, Philosophical Magazine 2, 557-72.

[15] Pitt, D., 2007. Modeling the Claim Duration of Income Protection Insurance Policyholders Using Parametric Mixture Models, Annals of Actuarial Science, 2(1),1-24.

[16] Rapach D.E., Strauss J.K., and Zhou G., 2010a.Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. Review of Financial Studies 23: 821C86.2

[17] Senensky B. and Polon J., 2004. Predicting Return to Work with Data Mining, Claim Analytics.

[18] Pearson, K., 1901 On lines and planes of closest fit to a system of points in space, Philosophical Magazine 2, 557-72.

[19] Pitt, D., 2007. Modeling the Claim Duration of Income Protection Insurance Policyholders Using Parametric Mixture Models, Annals of Actuarial Science, 2(1),1-24.

[20] Stock, J.H. and Watson, M.W., 2004. Combination forecasts of output growth in a seven-country data set, Journal of Forecasting 23, 405-430a.

[21] Wilkinson, L. and Dallal, G.E., 1981. Tests of significance in forward selection regression with an F-to enter stopping rule, Technometrics, 23, 377C380.

[22] 1991 Continuous Mortality Investigation Report No 12, UK, CMIB, Institute and Faculty of Actuaries.

[23] 1997 The Institute of Actuaries of Australia Report of the Disability Committee, Transactions of the Institute of Actuaries of Australia, 489-576.