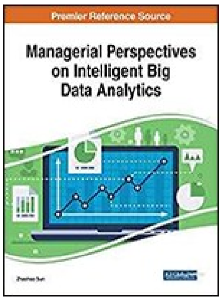

Managerial Perspectives on Intelligent Big Data Analytics

by Zhaohao Sun

IGI Global. (c) 2019. Copying Prohibited.

Chapter 10: Credit Rating Forecasting Using Machine Learning Techniques

Mark Wallis,
*Bond University,
Australia*

Kuldeep Kumar,
*Bond University,
Australia*

Adrian Gepp,
*Bond University,
Australia*

ABSTRACT

Credit ratings are an important metric for business managers and a contributor to economic growth. Forecasting such ratings might be a suitable application of big data analytics. As machine learning is one of the foundations of intelligent big data analytics, this chapter presents a comparative analysis of traditional statistical models and popular machine learning models for the prediction of Moody's long-term corporate debt ratings. Machine learning techniques such as artificial neural networks, support vector machines, and random forests generally outperformed their traditional counterparts in terms of both overall accuracy and the Kappa statistic. The parametric models may be hindered by missing variables and restrictive assumptions about the underlying distributions in the data. This chapter reveals the relative effectiveness of non-parametric big data analytics to model a complex process that frequently arises in business, specifically determining credit ratings.

INTRODUCTION

The notion of credit rating has been present in financial markets since 1860, where H.V. Poor began publishing financial statistics about railroad companies to attract public investments (Standard & Poor's, 2016). After this development, in 1909 J. Moody, founder of Moody's Investors Service, expanded on this idea by classifying these statistics into categories represented by letters of the alphabet. This methodology was mostly used on railway bonds. After the Great Depression in the late 1930's, the bond rating system became institutionalized in the United States. The repetitive nature of strong markets followed by crashes increased the need for a measure of risk and uncertainty for investors. Nowadays, 100 percent of all commercial papers and 99 percent of corporate bonds have been rated by at least one credit rating agency in the United States. These credit rating agencies have expanded across the globe to aid the needs of investors and corporate borrowers.

Credit ratings for companies have evolved to become an integral source of information for the financial sector. This information has a range of financial and economic benefits to society. These benefits can be categorized into three groups: benefits to investors, the company and the economy. The investors benefit from this information because it is a convenient and cost-effective source of information that allows for calculated risk. Furthermore, it encourages market confidence and entices retail investors to invest their savings into corporate securities and receive higher returns. For companies, credit ratings allow them to enter the market more confidently and raise funds at a lower cost. Companies may also use credit ratings as a means for brand repair or improvement. Lastly, with regard to the overall economy, consistent and accurate credit ratings fuel public investment in the corporate sector, which in-turn stimulates economic growth. These credit rating systems can facilitate the formation of public policy guidelines on institutional investors. They also play a vital role in investor protection by encouraging ethical behavior among corporate borrowers without putting a larger burden on the government.

Although they are not perfect, it is clear that credit ratings offer a plethora of benefits to society and are necessary to sustain strong economic growth and prosperity. These ratings are formed by incorporating a range of quantitative and qualitative variables that are gathered through public information and on-site research. However, these ratings take a substantial amount of labor and time to develop, making it a very costly process. This means it is difficult for management at many companies to afford regular credit rating updates. As a result, credit rating modelling has become a large area of research due to the economic and financial benefits associated with making credit ratings more efficient and cost-effective. With the expansion of machine learning and big data analytics over the past decade, there has been an influx of credit rating models in academic literature. As machine learning is one of the foundations of intelligent big data analytics, this chapter presents a comparative analysis of both traditional statistical models and popular machine learning models for the prediction of Moody's long term corporate debt ratings for top companies in the United States.

Moody's Rating System

Moody's, alongside Standard & Poor's and Fitch Group, is one of the three largest credit rating agencies in the world. The agencies all provide international finance research on bonds that are issued by both government and commercial entities. Moody's focuses on rating a borrower's creditworthiness based on a range of factors and rating scales that are designed to estimate the expected loss suffered by an investor in the event of a default and the probability of that event occurring. These rating systems are universally comparable, meaning they can be compared across different currencies, industries and countries. Moody's provide eight main categories of credit ratings (Moody's Investor service, 2017):

1. Moody's Long-Term Ratings
2. Moody's Short-Term Ratings
3. Moody's Bank Deposit Ratings
4. Moody's Bank Financial Strength Ratings
5. Moody's Mutual Fund Ratings
6. Moody's Insurance financial strength Ratings
7. Moody's issuer Ratings
8. Moody's management quality ratings for US affordable housing provider and National Scale Ratings

For the purpose of this chapter, the main focus will be placed on Moody's Long-Term Ratings. Moody's Long-Term Ratings are assigned to a range of fixed income instruments including bonds, debentures and preferred stocks. This rating system is a reflection of two major areas of the company. First, it reflects the credit risk of the company and how likely an issuer of debt is to meet its obligation. Secondly, it reflects the indenture protection, which represents the level of legal protection of the security. Some factors that may be included in the determination of this rating may include the seniority of the bond, negative pledge clauses and guarantees.

The following ratings are possible with Moody's Long-Term Ratings:

1. **Investment-Grade:** Aaa, Aa1, Aa2, Aa3, A1, A2, A3, Baa1, Baa2, Baa3, and
2. **Speculative-Grade:** Ba1, Ba2, Ba3, B1, B2, B3, Caa1, Caa2, Caa3, Ca and C.

In this chapter, these ratings are grouped into six categories as shown in Table 1, which is consistent with the earlier work of Kumar and Bhattacharya (2006). This categorization was used to ensure that a broad range of companies and credit ratings would be analyzed.

BACKGROUND

This section presents a brief review of reviewing the academic literature on credit rating forecasting with a focus on modelling techniques used to predict long-term credit ratings of corporate debt.

Studies about modelling credit rating have increased substantially over the past decade, likely because of the Global Financial Crisis and the realized importance of credit rating agencies in financial markets. Prior research identifies two main benefits gained from modelling credit ratings. First, it is very costly to employ a credit rating agency to provide long-term debt ratings more frequently than once a year, particularly for smaller companies who wish to improve their financial image in the market. Companies with good credit ratings enter the market with higher confidence and can raise funds at a cheaper rate (Kumar & Bhattacharya, 2006). On the other hand, for lending institutions and investors who create debt portfolios, it is important to have a regular indication of the riskiness of that portfolio. This also encourages economic growth through investment in the corporate sector. As a result, credit rating forecasting has become a popular area of study.

Table 1: Categories of Moody's ratings used in this chapter

Codes	Categories	Moody's Ratings
X1	High Grade	Aaa, Aa1, Aa2, Aa3
X2	Investment Grade	A1, A2, A3
X3	Upper Medium Grade	Baa1, Baa2, Baa3
X4	Medium Grade	Ba1, Ba2, Ba3
X5	Lower Medium Grade	B1, B2, B3
X6	Speculative Grade	Caa1, Caa2, Caa3, Ca and C

The credit rating forecasting literature is split into research using traditional statistical methods, machine learning techniques and ensemble techniques. The traditional statistical techniques used include logistic regression (Stepanova & Thomas, 2001; Steenackers & Goovaerts, 1989), linear discriminant analysis (Kumar & Bhattacharya, 2006; Khemakhem & Boujelbene, 2015) and Bayesian networks (Hajek, Olej, & Prochazka, 2016). These techniques were found to produce an average accuracy range of only 60 to 70% using financial statements ratios and other financial statement data. There was a consensus that accuracy could not be meaningfully increased further without gathering qualitative information such as that used by Moody's. As an extension to this hypothesis, Hajek, Olej & Prochazka (2016) explored the effect that news sentiment has on credit ratings. It was found that positive sentiment in the company's reports and negative words in news articles were statistically significant in predicting credit ratings.

Consistent with other areas of business modelling, there has been increased popularity in the application of non-parametric approaches to credit rating forecasting. The idea is to capture the non-linear relationships between variables and increase the predictive accuracy of the models without incorporating the qualitative variables used by Moody's. The most common techniques used include decision trees (e.g. Yobas & Crook, 2000) and k-nearest neighbors (KNN) (e.g. Henley & Hand, 1997). The performance of decision trees was found to outperform KNN when evaluated on imbalanced datasets with either a majority of low credit ratings or majority high credit ratings (Abdou & Pointon, 2011). However, there are often discrepancies in conclusions about the relative performance of these techniques when data sets of varying degrees of imbalance (skewness) (Abdou & Pointon, 2011; Henley & Hand, 1997). Overall, it is recognized that non-parametric approaches perform better than traditional statistical approaches (Abdou & Pointon, 2011).

Kernel classifiers and other modern machine learning techniques have also been applied to credit rating forecasting, encouraged by the success of other non-parametric approaches. A broad range of techniques applied and the most successful include Artificial Neural Networks (ANNs) (Kumar & Haynes, 2003; Kumar & Bhattacharya, 2006), Support Vector Machines (SVM) (Cristianini & Scholkopf, 2002) and a Gaussian Process Classifier (GPC) (Shian-Chang, 2011). Overall, SVMs have had the most success. SVMs are a popular model in this field because it is believed that the formulation of the SVM embodies the structural risk minimization principle. This means that the SVM produces an optimal trade-off between complexity and the empirical accuracy (Sewell, 2008). Despite these attractive features, many practitioners believe that SVMs ability to handle sparse data has potentially been overstated. For example, it has been shown that SVMs are not always able to construct parsimonious models in system identification and financial forecasting (Huang, Chuang, Wu, & Lai, 2010). It is also argued that SVM's underperform in their predictions because of the big data nature of the financial data used in credit rating forecasting – the curse of dimensionality (Huang et al., 2010).

This weakness in SVMs has sparked the use of many probability kernel classifiers such as the Gaussian process based classifiers (Girolami & Rogers, 2006). This technique has been prominent in statistics for decades and has outperformed ANNs on smaller datasets (Lilley & Frean, 2005). Huang (2011) explored the use of Gaussian processes in predicting credit ratings compared to using SVMs for prediction. GPCs were found to outperform conventional SVMs, even when enhanced by true selection and dimensionality reduction schemes, as tested on the Taiwanese banking sector (Shian-Chang, 2011). This is due to GPCs robustness when dealing with the high dimensionality of data using a fast-variational Bayesian algorithm proposed by Girolami & Rogers (2006) to reduce the computational loading of predictions.

Other modern research about modelling credit ratings uses ensemble techniques such as Random Forest (Wu & Wu, 2016) and Gradient Boosted Machines (GBM) (Abdou & Pointon, 2011). The resulting models are more difficult to interpret and some deem them to be similar to *black box* models. Nevertheless, they have produced high predictive performance and have a robust variable importance feature that partially explains the relationships between credit rating output and the input predictor variables (Imad, 2017).

Overall, the literature consists of a range of modelling techniques used to forecast long-term credit ratings. These techniques all have their own strengths and limitations and have all been tested on different data with diverse economic characteristics and distributions of credit ratings. There have been a range of papers claiming to produce the most accurate modelling technique; however, there has been a lack of comparisons between traditional statistical, non-parametric, ensembles and machine learning techniques on the same data set. Without such a comparison there is no solid evidence to suggest what technique results in the highest accuracy. The remainder of this chapter presents the findings from such a comparison.

DATA AND MODELLING TECHNIQUES

Data

Moody's Investor Service typically model credit ratings based on certain financial ratios. Financial ratios are often deemed to be good indicators of the company's financial health and security of the company (Ganeshalingam & Kumar, 2001). However, Moody's deems that the variables used to formulate these ratings are part of the company's intellectual property and so do not disclose these inputs (Moody's Investor service, 2017). As a result of this, financial ratios and other variables that are important in predicting the profitability, liquidity, and capital gearing of the company were gathered. The following 27 variables were considered when creating and formulating the credit rating models:

1. Operating Margin,
2. Pre-tax Margin,
3. Return on Invested Capital,
4. Return on Assets,
5. Current Ratio,
6. Quick Ratio,
7. Current Asset to Total Assets,
8. Operating Income to Net Sales,
9. Retained Earnings to Total Assets,
10. Accounts Receivable to Sales,
11. Inventory to Sales,
12. Sales to Total Assets,
13. Net Fixed Assets to Total Assets,
14. Long-Term Debt to Total Assets,
15. Total Liabilities to Total Liabilities and Equity,
16. Number of Employees,
17. Disposal of Fixed Assets,
18. Best Sales,
19. Total Assets,
20. Inventory to Current Assets,
21. Total Debt to Total Equity,
22. Total Debt to Total Capital Expenditure,
23. Cash Ratio,
24. Cash to Total Assets,
25. Asset Turnover,
26. Equity to Total Capital, and
27. Equity to Total Asset.

The above data were collected for 308 companies of the United States S&P 500 from Bloomberg with an examined time-period of January 2016 to November 2017. The companies were chosen such that Moody's Long-term Debt Ratings were available within the past five years, which resulted in useful data set for modelling. This data set also spans a range of industries and sectors, making it adequate to conduct a broad comparative evaluation.

Parametric Modelling Techniques

As mentioned in the literature review, credit rating modelling was initially conducted using parametric models such as multinomial logistic regression and linear discriminant analysis. These models are the best possible modelling techniques if the assumptions of the underlying data are satisfied. This section discusses the parametric models evaluated in this chapter.

Multinomial Logistic Regression

There are three main assumptions behind Multinomial Logistic Regression (MLR): observations are independent of one another, outcomes follow a categorical distribution derived from the covariates via a link function and there is a linear relationship between the covariates and the link-function-transformed outcome (Miyamoto, 2014). MLR is an extension from logistic regression in the sense that it produces a probability that an outcome belongs to a particular category. The link function used remains as logit, which is the logarithm of the odds – this correctly restricts the probability between 0 and 1. Other link functions are possible. For example, the probit model uses the inverse normal distribution function.

Linear and Regularized Discriminant Analysis

Observations are still assumed to be independent of one another by Linear Discriminant Analysis (LDA). In addition, data categories are assumed to be normally distributed (or at least symmetric), the covariance matrix is assumed to be the same for all categories and model accuracy is multicollinearity problems. LDA looks for linear combinations of variables that explain the data, where it focuses on attempting to model the differences between the categories of data and makes use of Bayes Theorem to estimate the probability of the output category given each input. Quadratic Discriminant Analysis (QDA) relaxes the assumption of a common covariance matrix. An extension to these modelling techniques is Regularized Discriminant Analysis (RDA) (Guo, Hastie, & Tibshirani, 2007), which is a weighted average of LDA and QDA. It introduces regularization in the modelling process to empirically determine a balance between a common covariance structure (LDA) and different covariance structures for each category (QDA). That is, RDA will automatically become more weighted towards QDA if the covariance matrix is not common to all categories. This model is the most statistically powerful and efficient technique if the data is a multivariate normal and homoscedasticity is present. In this chapter, both LDA and RDA will be evaluated.

Non-Parametric and Machine Learning Modelling Techniques

It is often argued that, particularly in smaller companies, there are not enough companies seeking financing to gain a full understanding of the population distribution of credit rating data. Therefore, it is not conclusive that the credit rating data meets the assumptions of parametric models (Guotai & Zhipeng, 2017). In fact, after a comprehensive review of such parametric assumptions, Elliot & Kennedy (1988) found that these assumptions often do not hold true for financial data. As a result, the following non-parametric modelling techniques are used in this chapter.

Artificial Neural Networks

As mentioned, Artificial Neural Networks (ANNs) have frequently been used to model credit rating. ANNs are soft computing techniques that are based on the neural structure of human brains. ANNs are able to model complexities that traditional quantitative methods used in finance and economics cannot due to the complexity in translating the system into precise functions. ANNs consist of three main layers: input, hidden and output, whereby more hidden layers are used to model more complex relationships in the data. Initially the neural connections within and between layers are set with random weights and then the model adjusts the weights during the learning process. If the prediction is correct, the ANN adjusts the weights in a positive way, whereas they are adjusted in a negative way if the outcome is negative. This modelling technique is useful when the underlying model structure is unknown, as the model has the ability to learn a wide variety of patterns from the data. ANNs have been shown to be effective at classification of groups and short-term predictions. They are also robust in the sense that they deal well with missing data and correlations between input variables (multicollinearity). However, ANNs are seen to be a *black box* technique because it is difficult to understand why the model makes the predictions it does. ANNs have also been known to over-fit the data used for training, which results in low accuracy for future predictions, particularly long-term predictions. Chapter 6 of Negnevitsky (2011) provides a more detailed introduction into ANNs.

Support Vector-Machines

Support Vector Machines (SVMs) are another type of supervised machine learning technique. They have increased in popularity in the literature in recent years (see Provost and Fawcett (2013, p. 89-94) for a brief introduction to SVMs). The SVM produces hyperplanes that divide data into the different categories. The hyperplanes are chosen such that they maximize the distance (margin) between the nearest data points of different categories. If data cannot be linearly separated, a kernel can be added to transform the data to assist separation by the SVM. This method often works effectively when there is a large number of variables (high dimensionality) and with low sample data. However, SVMs take a large amount of time to train and do not deal well with noisy data, which is data with inaccuracies.

Gaussian Process Classifier

Another non-parametric method that has been applied to credit rating forecasting is the Gaussian Process Classifier (GPC). This process is founded upon a Bayesian methodology. It assumes a prior distribution on the probability densities using the underlying mean and covariance of each variable (Shian-Chang, 2011). It assumes that each variable is drawn from a Gaussian distribution with the respective mean and variance and observing new elements creates a posterior distribution. The data is transformed using a squared-exponential kernel and parametrized using two parameters- sigma and L. Sigma (or L) dictates the height (or length) of the distribution that a point can be drawn from without being classified as an outlier. Predictions are formed by drawing from these collective underlying distributions and categorized into its appropriate category (Girolami & Rogers, 2006).

Random Forest (Ensemble)

Random Forest (Breiman, 2001) is an ensemble technique that has been applied to credit rating data with success in recent years. Random Forest utilizes multiple decision trees and bootstrapping to estimate an outcome. These decision trees are formed by recursively splitting data in two. The aim is that each one of the resulting number of groups identifies with a single category; it is common for multiple groups to be assigned the same category. One of the features of Random Forests is that only a random subset of variables is considered at each split. The other main feature is that each tree is grown to its maximum size, which means individually they are overly-complex, over-fit and will be poor future predictors. However, the accuracy is usually greatly improves then the individual predictions are combined. The way the combination is done is that each individual tree makes a separate prediction and then the category with the most votes (predictions) is assigned. This modelling technique is robust in large datasets with a large number of variables (high dimensionality), deals well with missing data and noise being present in the data, and has methods for adjusting to imbalanced data.

Gradient Boosted Machines (Ensemble)

Boosted machines combine multiple weak models together to form an accurate model. In the case of decision trees, a Gradient Boosted Machine (GBM) (Friedman, 2001; Friedman, 2002) first estimates an overly-simple decision tree (with shallow depth). The GBM then slightly improves the model based on the its prediction error. This process of gradual improvement is then repeated a large number of times to ideally achieve a substantially higher accuracy. This ensemble technique has been less popular than ANNs, but it has often produced higher accuracy (Imad, 2017).

PROCEDURE AND RESULTS

This section outlines the process behind selecting the parameters for the different models and an analysis of the predictive accuracy of each model. It also outlines the most important variables for each model.

The two measures for evaluating accuracy are the overall accuracy (percentage of correct classifications) and Cohen's Kappa statistic (Kappa). Kappa is an accuracy measure that considers both the observed accuracy of the model and the expected accuracy. It is calculated as

$$\frac{Observed\,Accuracy - Expected\,Accuracy}{1 - Observed\,Accuracy}.$$

The inclusion of the expected accuracy allows the statistic to adjust to imbalanced data sets that do not have an equal number of data points in each category. This is particularly important given that credit rating forecasting often involves imbalanced data, as mentioned earlier in the Background. A more detailed explanation of Kappa is provided in the Appendix.

In the literature, the predictive performance of models is commonly compared using hold-out or test data. This process involves training models on a random subset of the data (typically 70%) and then measuring its predictive performance on a hold-out testing set (30%); thus, obtaining a real-world estimate of model performance using new data unseen by the model. However, the resulting single estimate of accuracy can be greatly influenced by precise data split that occurred. Although very big data sets mitigate this problem, it is possible that very different results could be obtained if a different split were randomly chosen. To avoid this issue, the analysis presented in this chapter uses repeated 10-fold cross-validation, where the number 10 is a standard through modelling literature. 10-fold cross validation involves randomly splitting the data into 10 approximately equal partitions, then training the model on 9 of the data subsets and evaluating the model on the final subset. This train-evaluate process is repeated for all 10 possibilities for the subset that is used for evaluation. Further, this whole 10-fold process is then repeated 5 times, each time with a different randomly chosen 10 subsets. This results in a total of 50 accuracy measures per model. These 50 figures can be averaged to obtain a more stable estimate of model performance. Moreover, the 50 different estimates allows for the variability in model performance to be assessed and confidence intervals for both to be estimated for both overall accuracy and the Kappa statistic. It is widely understood that compared to single point estimates, confidence intervals can greatly enhance the interpretability from a managerial perspective.

Importantly, the repeated cross-validation process is common to all models. That means that the data splits are identical for all models, which enables comparison between models to be fair and valid.

Multinomial Logistic Regression

The top 20 most important variables are shown in Figure 1. This ranking was determined by determining the Area Under the receiver operating characteristic Curve (AUC) for each category pair (i.e. High Grade vs Investment Grade, Investment Grade vs Upper Medium Grade and so on). For a specific category, the maximum AUC for all the relevant pairs is used as the variable importance measure (Abdou & Pointon, 2011). The top three most important variables are long-term debt, retained earnings to assets and return on assets.

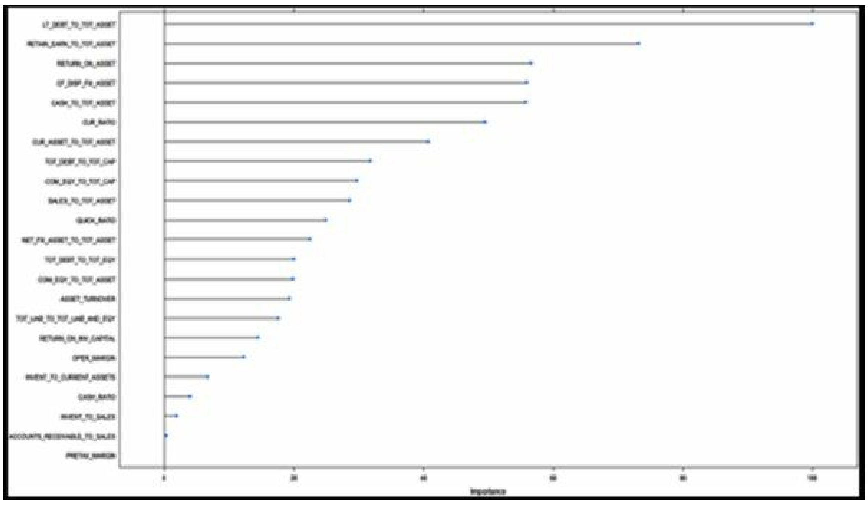

Figure 1: Variable importance ranking according to multinomial logistic regression

Linear and Regularized Discriminant Analysis (LDA and RDA)

The choice of model parameters can have a large influence on model accuracy; consequently, the main parameters for each model are empirically optimized using cross-validation. As with MLR, LDA does not require any parameters to be set. However, as RDA is a weighted average of LDA and QDA, the relative weights for each are determined by maximizing accuracy using cross-validation. Interestingly, the resulting RDA model weighted LDA at 100% and QDA at 0%, indicating that the same covariance structure is common to all rating categories.

In this case, LDA and RDA have the same variable importance. Similar to most machine learning techniques, discriminant analysis considers variable importance for each rating category separately. Figure 2 shows these rankings. Return on Assets is the most important predictor for all ratings except for X3 and X4, where the most important predictor is Total Assets. It is interesting to note that discriminant analysis ranks pre-tax margin as the second most important variable out of all the variables, whereas MLR ranked it last. The most important variable in multinomial logistic regression was Long-term debt to total assets and this was the 8th most important variable in LDA and RDA.

Artificial Neural Network (ANN)

The main parameters set for the ANN include the number of hidden layers and the weight decay. Weight decay is a weight update rule that causes the hidden layer weights to exponentially decay to zero if no other update is issued. Up to three hidden layers were trialed and again cross-validation was used to determine the optimum. The training tolerance was also set to 0.01 to ensure a balance between accuracy and over-fitting. The optimal model was found to be an ANN with 2 hidden layers and a weight decay of 0.1.

As shown in Figure 3, the variable importance rankings appear to be similar to RDA across all rating categories except for Debt to Assets, Number of Employees and Best Sales. The importance across the categories shift in order compared to RDA.

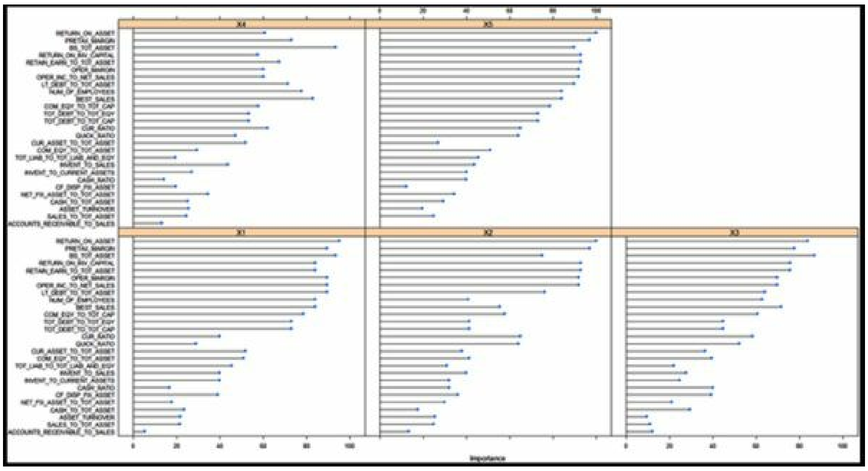

Figure 2: Variable importance ranking according to discriminant analysis

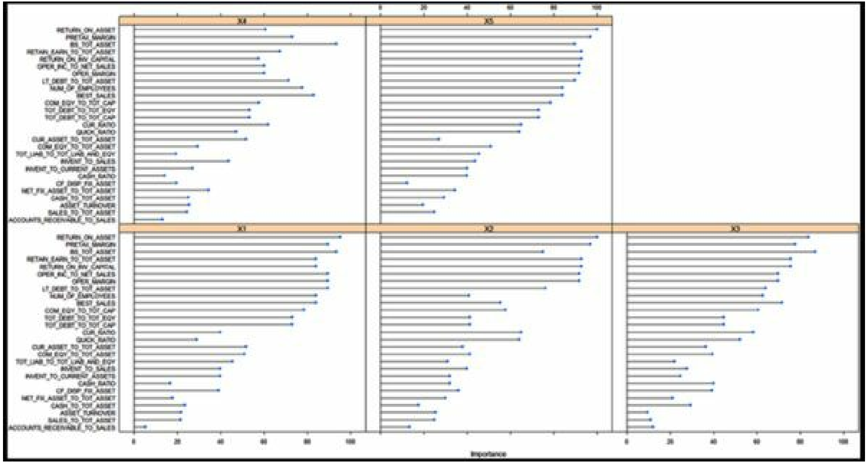

Figure 3: Variable importance ranking according to artificial neural network

Support Vector Machine (SVM)

SVMs can be tuned using two variables in the freely available R programming language. These variables are Sigma and Cost and together they control the trade-off between the accuracy on the training data and the risk of over-fitting (resulting in poor future predictions). The settings that yielded the highest cross-validated accuracy were a Sigma of 0.03125 and Cost of 5.

The variable importance for the SVM is calculated by computing the AUC. As shown by 8, the variable importance rankings of the SVM have remained relatively similar to that of the ANN. There is only one material changes that occur in the variable importance rankings. First, Return on Invested Capital overtook Retained Earnings to Total Assets for 4th most important variable.

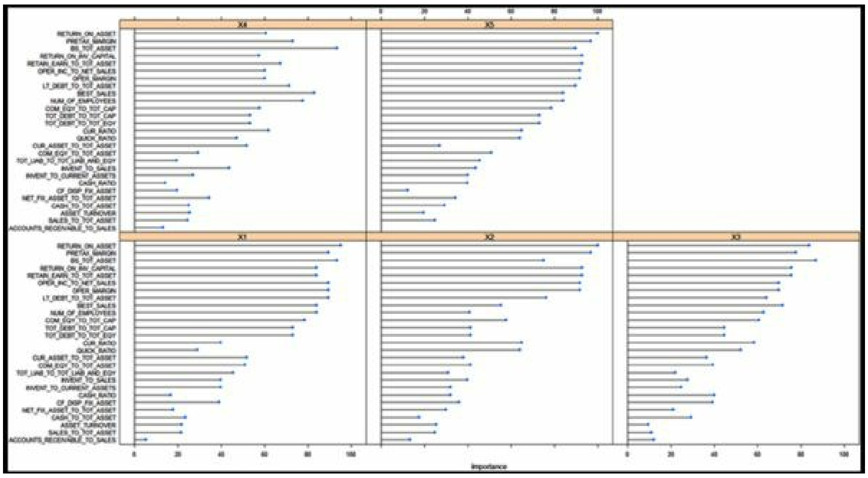

Figure 4: Variable importance ranking according to support vector machine

Gaussian Process Classifier (GPC)

The GPC in the R programming language can be tuned using one parameter when modelled using the popular Radial Basis kernel function-sigma. This sigma is a hyper-parameter for the kernel; it indicates the width of the Radial Basis of the kernel function and the Laplacian kernel (Shian-Chang, 2011). In essence, it represents how far away a data point needs to be before it is deemed an outlier. A higher sigma may lead to a more accurate model, but it also makes the model susceptible to over-fitting. This is essentially the same trade-off between accuracy and over-fitting. The optimal model was found to have a sigma of 0.135, chosen based on cross-validated accuracy.

The variable importance for a GPC is generated according to the technique specified by Linkletter, Bingham, Hengartner, Higdon, & Ye (2006). This methodology uses a Bayesian approach whereby it analyses the effects on the posterior distribution. As shown in Figure 5, the variable importance is essentially the same as for the SVM.

Random Forest (RF)

RF is controlled by two main parameters: the number of trees in the ensemble and the number variables to be randomly selected at each split point for each tree. 500 trees were grown based on the size of the data set. The number of variables to randomly choose is often chosen as the square root of the number of variables ($\sqrt{27} \approx 5$). However, it is prudent to empirically verify the optimum setting using cross-validation. This resulted in a setting of four, similar to the square root heuristic that suggested five.

The variable importance of the RF model is ranked using the standard metric: mean decrease in node impurity. This allows an overall ranking to be calculated, rather than a different ranking for each category. As shown in Figure 6, the variable importance is notably different to all other models. The most important variable is Best Sales, which is typically ranked around 4th in SVM and GPC and did not even make it into the top 20 most important variables for MLR. This difference is likely because unlike the other models, Random Forest is an ensemble of multiple models and so its predictions can arguably be considered more stable.

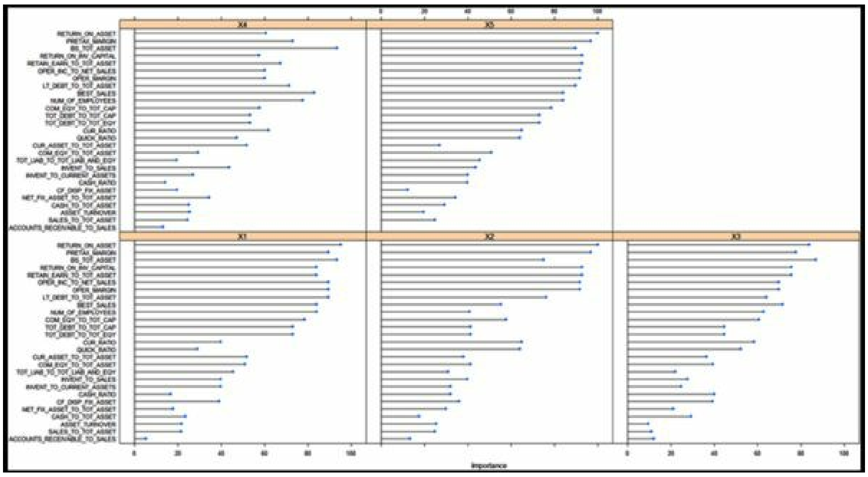

Figure 5: Variable importance ranking according to a Gaussian process classifier

Gradient Boosted Machines (GBMs)

GBMs can be parametrized using three main variables: the number of trees, the shrinkage (or learning) factor and the size of each individual tree. The number of trees represents the amount of weak learning predictors that will be grown in the simulation and the shrinkage factor represents the learning rate of the algorithm. The learning rate determines the rate at which it shrinks the impact of incorrect predictors. Using cross-validation, the optimal settings are 1000 trees, shrinkage of 0.005 and a minimum of 30 nodes per tree.

The variable importance method used is similar to the method used for Random Forest. The importance is measured based on the mean decrease in impurity. As shown in Figure 7, the most important variable is Return on Assets followed by Best Sales. This is similar to Random Forest, the other ensemble model, but different to all the other models.

Comparison

Table 2 shows the average results for all models, while Figure 8 shows the distribution of overall accuracy and Kappa statistic results. Based on Figure 8, an overall ranking is obtained and is shown in Table 2. All modelling techniques appeared to produce relatively similar mean average accuracy, but the top three techniques produced the largest improvement above guessing (Kappa statistic). The focus of this chapter is the comparison between models.

In general, the lowest ranked models are the parametric techniques. This makes sense as these modelling techniques make a range of assumptions that are likely not satisfied by financial data. The multinomial logistic regression is the poorest performing modelling technique in terms of average accuracy and the second worst in terms of Kappa, with a mean accuracy of 59.6% and a mean Kappa statistic of 22.6%. It is interesting to note that RDA performs worse than LDA. The two models are similar, but the RDA includes a gamma penalty parameter (optimized by cross-validation) that appears to reduce the accuracy of the model.

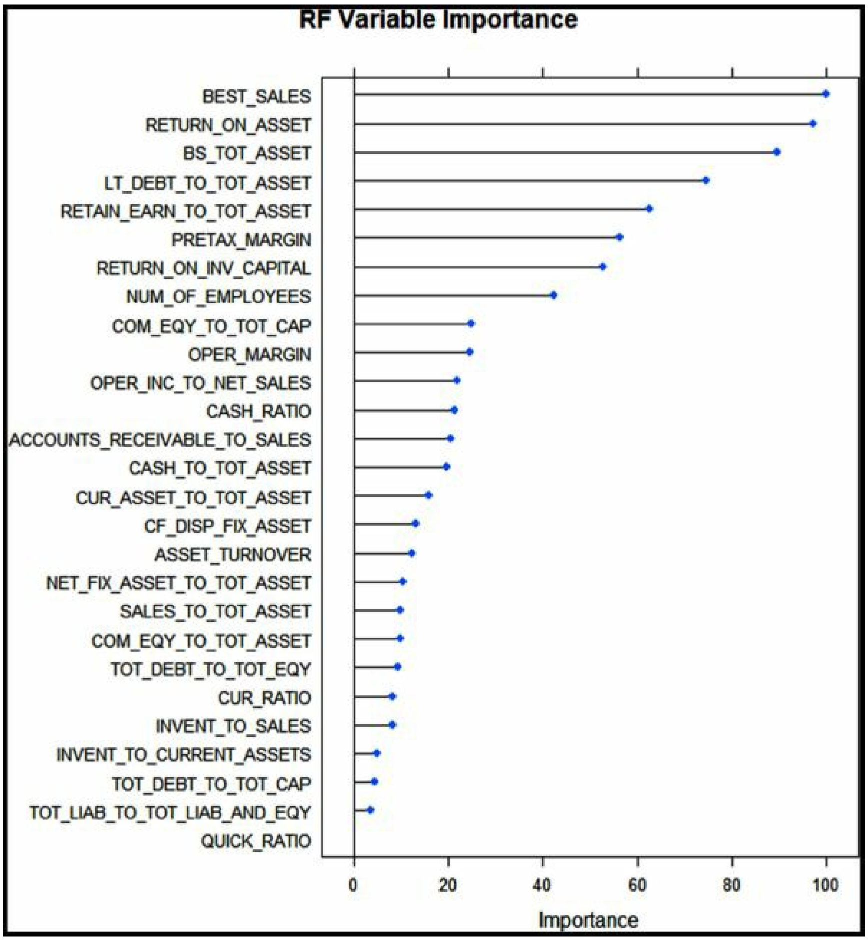

Figure 6: Variable importance ranking according to Random Forest

The ANN performed the best in terms of the Kappa statistic, suggesting it had the largest improvement compared to guessing the credit rating by chance. However, the overall accuracy was only the 6th best. The GPC performed relatively poorly compared to what was reported in Shian-Chang (2011), performing the worst in terms of Kappa and 5th in overall accuracy. The best overall performer was Random Forest, producing the highest overall accuracy and second highest Kappa statistic. This was followed closely by SVM, which produced the second highest overall accuracy and the third highest Kappa.

CONCLUSION

This chapter presented a comparative study of some of the most popular modelling techniques used to model credit ratings. These models were evaluated on 308 of the S&P 500 companies to predict the Moody's long-term credit rating. It was found that non-parametric techniques usually outperformed parametric techniques, likely because the underlying assumptions were not met by financial data. The top three performing modelling techniques were Random Forest, Artificial Neural Networks and Support Vector Machines. They produced models with an average accuracy of 64.6%, 63.6% and 60.1% respectively.

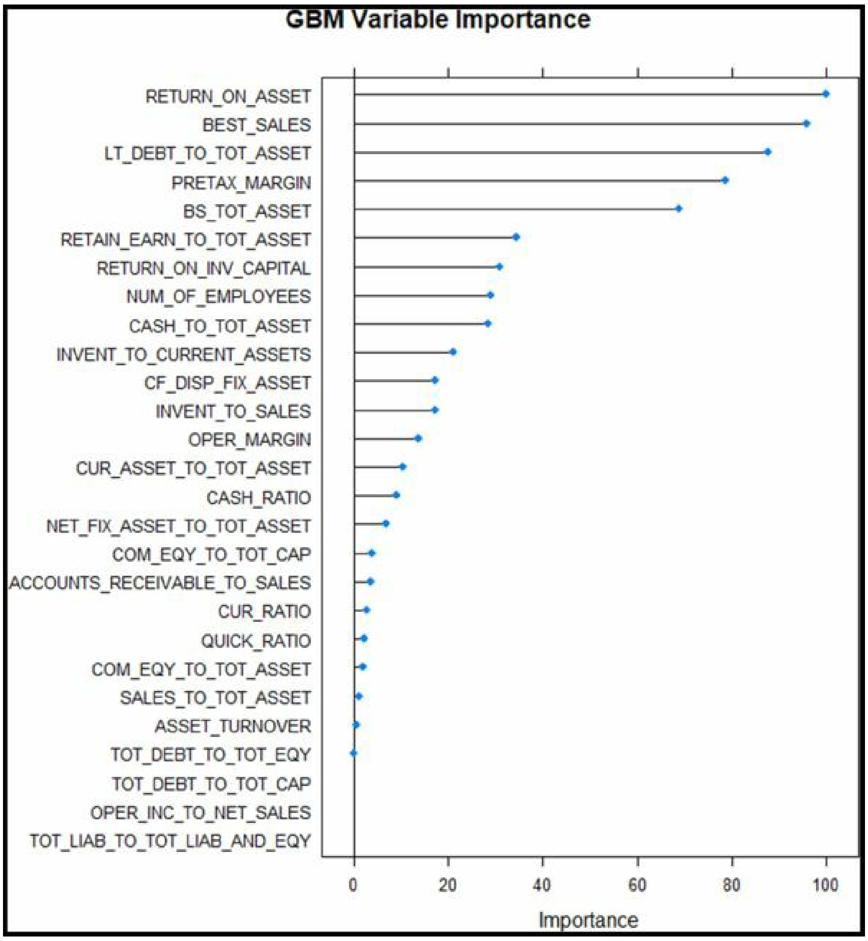

Figure 7: Variable importance ranking according to Gradient Boosted Machine

Table 2: Comparison of model accuracy

Method	Overall Ranking	Average Accuracy	Average Kappa
RF	1	64.6%	31.3%
SVM	2	63.6%	30.6%
ANN	3	60.8%	31.5%
GBM	4	62.3%	27.4%
LDA	5	61.7%	28.8%
RDA	6	60.4%	25.3%
MLR	7	59.6%	22.6%
GPC	8	61.6%	20.6%

Credit rating forecasting is an example of an important business process that is better modelled by flexible machine learning models that, unlike traditional parametric techniques, do not make restrictive assumptions about the data. The machine learning techniques are also better able to handle collinearity between variables in the model that can cause serious problems in traditional models. Overall, as machine learning is a key component of big data analytics, credit rating forecasting is a clear example of the value and importance of big data analytics in the future business world.

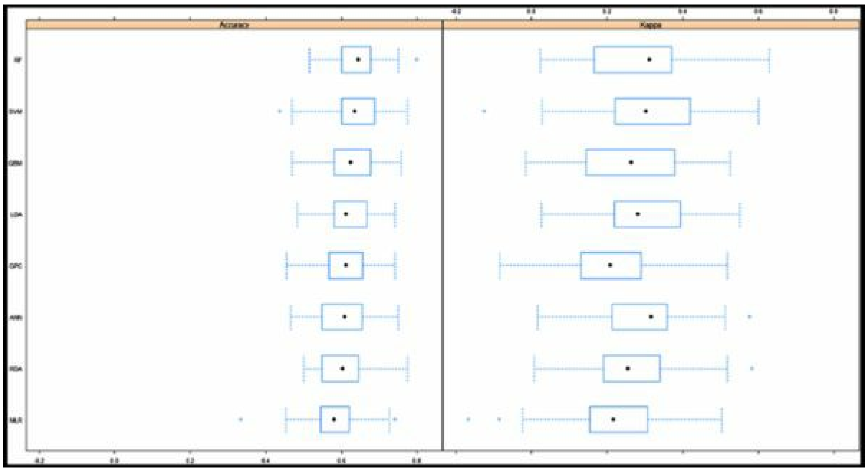

Figure 8: Box plot of model performance

An advantage of big data analytics is its adaptability to non-traditional data sources. One such example is automated sentiment analysis that determines whether there is positive or negative sentiment in text, such as annual reports or business news. Future research could incorporate the sentiment score output from big data analytics into the credit ratings models presented in this chapter. This has the potential to substantially increase accuracy given that Hajek et al. (2016) found credit ratings were affected by positive words in the relevant company's reports and negative words in associated news articles.

REFERENCES

Abdou, H., & Pointon, J. (2011). *Credit scoring, statistical techniques and evaluation criteria: A review of the literature*. *Intelligent Systems in Accounting, Finance & Management*, 18(2-3), 59–88. doi:10.1002/isaf.325

Breiman, L. (2001). *Random Forests*. *Machine Learning*, 45(1), 5–32. doi:10.1023/A:1010933404324

Cristianini, N., & Scholkopf, B. (2002). *Support Vector Machines and Kernel Methods: The New Generation of Learning Machines*. *AI Magazine*, 23(3), 31–41.

Elliot, J. A., & Kennedy, D. B. (1988). *Estimation and prediction of categorical models in accounting research*. *Journal of Accounting Literature*, 7, 202–242.

Friedman, J. H. (2001). *Greedy function approximation: A gradient boosting machine*. *Annals of Statistics*, 29(5), 1189–1232. doi:10.1214/aos/1013203451

Friedman, J. H. (2002). *Stochastic gradient boosting*. *Computational Statistics & Data Analysis*, 38(4), 367–378. doi:10.1016/S0167-9473(01)00065-2

Ganeshalingam, S., & Kumar, K. (2001). *Detection and prediction of financial distress*. *Managerial Finance*, 27(4), 45–55. doi:10.1108/03074350110767132

Girolami, M., & Rogers, S. (2006). *Variational Bayesian multinomial probit regression with Gaussian process priors*. *Neural Computation*, 18(8), 1790–1817. doi:10.1162/neco.2006.18.8.1790

Guo, Y., Hastie, T., & Tibshirani, R. (2007). *Regularized linear discriminant analysis and its application in microarrays*. *Biostatistics (Oxford, England)*, 8(1), 86–100. doi:10.1093/biostatistics/kxj035 PMID:16603682

Guotai, C., & Zhipeng, Z. (2017). *Multi Criteria Credit Rating Model for Small Enterprise Using a Nonparametric Method*. *Sustainability*, 9(10), 1834. doi:10.3390u9101834

Hajek, P., Olej, V., & Prochazka, O. (2016). *Predicting Corporate Credit Ratings Using Content Analysis of Annual Reports – A Naïve Bayesian Network Approach*. In S. Feuerriegel & D. Neumann (Eds.), *Enterprise Applications, Markets and Services in the Finance Industry* (pp 47-61). Springer.

Henley, W. E., & Hand, D. J. (1997). *Construction of a k-nearest neighbour credit-scoring system*. *IMA Journal of Management Mathematics*, 8(4), 305–321. doi:10.1093/imaman/8.4.305

Huang, S.-C. (2011). *Using Gaussian process based kernel classifiers for credit rating forecasting*. *Expert Systems with*

Applications, 38(7), 8607–8611. doi:10.1016/j.eswa.2011.01.064

Huang, S.-C., Chuang, P. J., Wu, C. F., & Lai, H. J. (2010). *Chaos-based support vector regressions for exchange rate forecasting*. *Expert Systems with Applications*, 37(12), 8590–8598. doi:10.1016/j.eswa.2010.06.001

Imad, B.-H. (2017). *Bayesian credit ratings: A random forest alternative approach*. *Communications in Statistics. Theory and Methods*, 46(15), 7289–7300. doi:10.1080/03610926.2016.1148730

Khemakhem, S., & Boujelbene, Y. (2015). *Credit Risk Prediction: A comparative study between discriminant analysis and the neural network approach*. *Accounting and Management Information Systems*, 14(1), 60–78.

Kumar, K., & Bhattacharya, S. (2006). *Artificial neural network vs linear discriminant analysis in credit ratings forecast*. *Review of Accounting and Finance*, 5(3), 216–227. doi:10.1108/14757700610686426

Kumar, K., & Haynes, J. D. (2003). *Forecasting credit ratings using an ANN and statistical techniques*. *International Journal of Business Studies*, 11(1), 91–108.

Lilley, M., & Frean, M. (2005) *Neural Networks: A Replacement for Gaussian Processes?* In *Intelligent Data Engineering and Automated Learning - IDEAL 2005* (pp. 195-212). Springer. doi:10.1007/11508069_26

Linkletter, C., Bingham, D., Hengartner, N., Higdon, D., & Ye, K. Q. (2006). *Variable selection for Gaussian process models in computer experiments*. *Technometrics*, 48(4), 478–490. doi:10.1198/004017006000000228

Miyamoto, M. (2014). *Credit risk assessment for a small bank by using a multinomial logistic regression model*. *International Journal of Finance and Accounting*, 3(5), 327–334.

Moody's Investor service. (2017). *Moody's rating system in brief*. Retrieved from *https://www.moodys.com/sites/products/ProductAttachments/ Moody's%20Rating%20System.pdf*

Negnevitsky, M. (2011). *Artificial intelligence: a guide to intelligent systems*. Harlow, UK: Addison Wesley/Pearson.

Provost, F., & Fawcett, T. (2013). *Data science for business*. Sebastopol, CA: O'Reilly Media.

Sewell, M. (2008). *Structural Risk Minimization*. Technical Report: Department of Computer Science, University College London. Retrieved from *http://www.svms.org/srm/*

Shian-Chang, H. (2011). *Using Gaussian process based kernel classifiers for credit rating forecasting*. *Expert Systems with Applications*, 38(7), 8607–8611. doi:10.1016/j.eswa.2011.01.064

Standard & Poor's. (2016, October 12). *Standard & Poor's History*. Retrieved from *https://www.isin.net/standard-poors/*

Steenackers, A., & Goovaerts, M. J. (1989). *A credit scoring model for personal loans*. *Insurance, Mathematics & Economics*, 8(1), 31–34. doi:10.1016/0167-6687(89)90044-9

Stepanova, M., & Thomas, L. C. (2001). *PHAB scores: Proportional hazards analysis*. *The Journal of the Operational Research Society*, 52(9), 1007–1016. doi:10.1057/palgrave.jors.2601189

Wu, H.-C., & Wu, Y.-T. (2016). *Evaluating credit rating prediction by using the KMV model and random forest*. *Kybernetes*, 45(10), 1637–1651. doi:10.1108/K-12-2014-0285

Yobas, M. B., & Crook, J. N. (2000). *Credit scoring using neural and evolutionary techniques*. *IMA Journal of Management Mathematics*, 11(2), 111–125. doi:10.1093/imaman/11.2.111

APPENDIX

Cohen's Kappa statistic is best explained using an example. The following example is related to 50 events, each which is either *Yes* or *No*. The goal of the model is to predict the *Yes/No* result of each event. The accuracy of the result is derived from the underlying truth of the variable. As shown by Table 3, out of the 50 events, 20 were correctly predicted *yes* and 15 were correctly predicted *no*. That means the observed accuracy is (20+15)/50=0.70 or 70%.

This accuracy however does not consider the expected accuracy based on the fact that, in truth, there are more *Yes* events than *No* events. Using the following equation, the Kappa statistic considers the expected chance of agreement between the truth and the model.

$$K = (P_o - P_e) / \left(1 - P_e\right),$$

where P_o represents the observed agreement and P_e represents the expected agreement. The observed agreement is the accuracy calculated in the example above; in this case, 70%. The expected accuracy is calculated by summing the probability that the truth and the model agree on yes and agree on no:

1. The model predicted *Yes* 25 (20+5) times and *No* 25 (10+15) times, which means the model predicts *Yes* 50% of the time;
2. In truth, *Yes* occurs 30 (20+10) times, which represents 60% (30/50);
3. Therefore, the probability that, at random, the model correctly predicts *Yes* is 0.5 × 0.6 = 0.3. The same calculation for *No* events is (1 - 0.5) × (1 - 0.6) = 0.2. This process is then continued if there are more than two categories. Finally, the sum of these two figures is calculated as the expected accuracy, which is in this case 0.3 + 0.2 = 0.5.

The Kappa can then be calculated, as

$$K = \frac{0.7 - 0.5}{1 - 0.5} = 0.4$$

Table 3: Visits to public libraries

		Truth	
		Yes	**No**
Model Prediction	**Yes**	20	5
	No	10	15