Predictive Models to Support Quoting of Fixed Fee Consulting Projects

Amy Cook, Paul Wu, Kerrie Mengersen*,[a]

[a]*School of Mathematical Sciences, George Street, Brisbane, QLD, 4000*

Abstract

A concise abstract is required. limit to 250 words. clearly state purpose of research, principal resutls and major conclusions. no refs

Engaging in loss making jobs for fixed fees is a major problem in consulting, particularly in the competitive construction industry. This thesis investigates whether machine learning techniques applied to a company's passively collected internal data could help avoid loss making jobs or help tactfully choose when to enforce stricter contracts. It was found that in a specific decision framework, a case study's profits could be improved 9% by declining approximately 4% of projects. Alternative decision frameworks are also proposed and evaluated. Algorithmic methods such as Logistic Regression, Random Forests, Boosted Trees, Naive Bayes, and Bayesian Networks were applied as well as blended combinations of these methods. A decision scenario which rejected projects above a sequence of tested thresholds was run in order to find the optimal threshold for profit improvements. The blended Logistic Regression model outperformed other methods and produced a 95% confidence interval of 6.5 - 11.5% profit improvements. The findings from this research have the potential to assist managers in reducing losses by highlighting risky projects and guiding project-based changes to fee structures.

Key words: consulting; machine learning; profitability; predictive model; construction industry; data mining,

Text based on elsarticle sample manuscript, see http://www.elsevier.com/author-schemas/latex-instructions#elsarticle

1. Introduction

Clearly state the research question and objectives of the work. Briefly provide any necessary background to frame the research question. Concisely summarize the major findings/results.

1 page

Intro - pick out bits from thesis

1.1 Problem motivation. for consulting businesses that solve complex problems, similar time estimation errors determine their financial well-being.
This thesis focuses on the risk taken by a consulting company when offering a fixed fee. Clients commonly seek fixed price quotes from several consultants before selecting one or negotiating further. Hence, quoted fees must be competitive. The way a consulting manager calculates a fixed price varies from industry to industry and even company to company. Typically a consulting manager has experience in the type of project they are quoting and, after reviewing the project details, can use a combination of intuition, comparison to past projects, and rules of thumb - amount of losses, sources A study across 1471 IT projects showed that 27% of projects ran over budget, and one in 6 of those projects were more than 200% over budget (Flyvbjerg 2011). A study on large-scale infrastructure projects over the past seventy years revealed that cost forecasts consistently underestimated the cost of rail projects by an average of 44.7%, the cost of bridge and tunnel projects by 33.8% and the cost of

*Corresponding Author
Email address: a21.cook@qut.edu.au (Amy Cook, Paul Wu, Kerrie Mengersen)

road projects by 20.4% (Flyvbjerg 2007). Smaller consulting companies in competitive industries, such as the construction industry, experience similar difficulties in forecasting their project costs. losses = low employee morale, elevated stress, low work quality to improve speed, possibly redundancies Lovallo and Kahneman (2003) delved into the psychology behind why executives so often severely underestimate costs of larger projects such as manufacturing plant construction, mergers and acquisitions, large infrastructure and software development. Their theory stemmed from Kahneman's work on decision-making that won him the Nobel Prize for economics in 2002. His research argued that a person's natural optimistic view of their own skills leads to consistent underestimation of the time and risks involved in a project. A manager optimistically sees challenges in a project as something that can be overcome by the team's high skill level, and downplays or ignores the risk of problems that are out of the team's control. all complex projects are at risk of encountering a multitude of problems that the manager could never foresee. Each problem has a low chance of occurring, but in combination the risk is much greater (Lovallo and Kahneman 2003).

Unfortunately, limited research is available that documents the methods industry uses for complex-projects. The construction cost estimation study by Akintoye and Fitzgerald (2000) surveyed 84 UK construction contractors - detailed analysis and experience based models A survey by Moores and Edwards (1992) of 54 software developing companies found that detailed project planning tools were used by most companies as opposed to cost estimation tools

Detailed analysis remains the most prevalent method for construction cost estimation despite the industry having a long history of projects running over time and budget using the same technique (Shane et al. 2009).

1.2 Case Study. This thesis focuses on a single case study consulting company in the construction industry. twelve years of passively collected data that described each project in terms of their clients, invoice history, employee hours and technical details. labour is the chief cost and the long traditional history of this industry mandates fixed price projects as the norm. 20% of projects are loss making. Fees calculated via detailed analysis A Customer Relationship Management software package (CRM) is currently employed to collect and store project data. A CRM is a popular type of software used by businesses to record client and project details as well as employee time sheet records. In the case study business, the CRM is available to all employees over the company intranet, and each employee completes daily time sheets allotting their hours to certain projects. Additionally, technical information is recorded against each project

- Employee time sheet hours with dates
- Other project costs (taxis, printing)
- Client information/characteristics
- Client identification code
- Invoiced amounts for each project and dates
- Employee costs
- Employee charge out rates
- Project description

2364 past projects are available. missing data due to how the data collection changed over time and some fields were not mandatory.

- aim **General Aim** Use statistical techniques to model the profitability of projects for consulting businesses using their internal CRM data. Research will focus on a case study Engineering consulting company that offers their expert advice (in the currency of time) to business clients. The project outcomes are intended to assist the business in predicting project profitability before engagement. Several statistical and machine learning techniques were tested, compared and refined.

Hypothesis 1 A statistical or machine learning model based on historical project data can predict the profitability of a new project with greater accuracy than a baseline predictor. A baseline predictor for this project is one that predicts the average of a numeric response variable for all cases or, if the response variable is categorical, randomly assigned categories for each case, where the proportion of assigned categories matches the true categorical proportions.

Hypothesis 2 The predictive model built from Hypothesis 1 can be shown to have a positive impact on the overall profit earned by the case study business. The overall profit is represented by the following equation:

- brief summary of major findings/results

2. Literature Review

Summary of Key Related Research This section should include a brief summary of key related research. Emphasis should be on demonstrating the foundation for the current investigation. Specifically, the goal is to clearly delineate a gap or missing link that the current research fills. Authors should avoid presenting a litany of past research and should focus on prior work necessary to demonstrate the existence of the research gap addressed in the manuscript.

1 page

2.1 Cost estimation in the Construction Industry and IT Industry. The bulk of research to date has been performed with project data in either the construction industry or software development. Research in the construction industry has primarily focused on predicting the final building cost of construction. Over the past few decades, significant research has been dedicated to creating predictive models that present the 'outside view' of a project. This was done by statistically comparing a new project to a collection of similar projects and their characteristics. It has been shown that these models improve cost estimation accuracy, however industry has not adopted them.

2.1.1 Construction Industry Elfaki, Alatawi, and Abushandi (2014) completed a review of cost estimation research from 2004 to 2014. They found that artificial Neural Networks and SVM's were the most common machine learning techniques. Neural Networks and SVM's received significant attention in the 1990's for their ability to accurately predict construction costs with limited detailed information (Shin 2015; Kim, An, and Kang 2004). Shin (2015) pioneered the application of Boosted Trees to cost estimation in construction projects. Often, Linear Regression was the only model assessed, without comparison to other methods such as Neural Networks, which first started appearing in literature in the 1990's (Kim, An, and Kang 2004) Some studies showed that Neural Networks outperform Regression, however other studies established they are approximately equal (Kim, An, and Kang 2004; Attalla and Hegazy 2003).

Variety of variables used: (Chan and Park 2005)(Akintoye and Fitzgerald 2000; Trost and Oberlender 2003; Pinto and Slevin 1988) the studies all generally reported their mean absolute error or a similar metric from the tested models, and results were generally positive. Furthermore, the existing studies collected data from many businesses for a single study

2.1.2 Software Industry In the software industry, the main component of cost is *effort* as opposed to the cost of building materials in the construction industry.

Similar to the construction industry, expert judgment or detailed analysis, is the most widely practiced method for effort estimation (Shepperd, Schofield, and Kitchenham 1996; Moløkken and Jørgensen 2003)

Multiple studies have shown that Neural Networks definitively outperform regression models in effort estimation, although Regression is the most popular method (Finnie, Wittig, and Desharnais 1997; Pai, McFall, and Subramanian 2013; Matson and Mellichamp 1993).

Interestingly, several studies have found that even if 15 or so variables are included, often only one variable contributes significantly to the model's accuracy: size (or funciton point) (Shepperd, Schofield, and Kitchenham 1996; Finnie, Wittig, and Desharnais 1997; Pai, McFall, and Subramanian 2013)

A study by Mendes and Kitchenham (2004) demonstrated using 67 web projects that cross-company models were significantly less accurate than a within company model (Shepperd, Schofield, and Kitchenham 1996) Again, Neural Networks were criticised for their inability to explain their results (Finnie, Wittig, and Desharnais 1997). The software industry is unique in that agile methods of delivery are changing the contractual approach of consulting (Badenfelt 2011). Software development is a relatively new field in comparison to construction, where the tradition of fixed fee structures is difficult to reinvent.

2.2 Methods used in other business applications. The review of methods applied to the construction and IT industry highlighted the use of Linear Regression, Neural Networks, SVM's and in one case Boosted Trees, however research on other business problems utilised a wider range of methods. These included Naive Bayes, Random Forests, and machine learned Bayesian Networks.

Although SVM's performed the best, as discussed, they are often complex and slow, requiring a great deal of memory (Kumar and Ravi 2007). Kumar and Ravi (2007) performed a detailed review of statistical and machine learning techniques that were applied over 37 years in the context of bankruptcy prediction in banks.

For example in financial credit scoring, a study by Brown and Mues (2012) found that Random Forests and Boosted Trees consistently outperformed Neural Networks in classification. Kumar and Ravi's (2007) review also assessed ensemble techniques, which refers to combinations of two completely different algorithms, and found they can often outperform individual methods.

An exception is Saradhi & Palshikar's (2011) study on employee churn, where 'churn' refers to the number of individuals moving out of a group within a certain time. Naive Bayes, Random forests (an ensemble decision tree method) and SVM's were built. determining the value of each employee in terms of the importance of the projects they were on and their monthly chargeability. This allowed them to rank employees identified as 'high risk of churn' by value and provided a clear ranking for manager's to act upon. This extension of the study provided a comprehensive framework for how business managers could adopt their findings to improve business operations. It was a valuable addition that is absent from most cost and effort estimation research.

2.3 Gaps. In comparison to previous studies, this project advances the body of work on cost estimation in three ways: the data set is larger than other studies and sourced from a single company's CRM database, ensemble tree methods and model blending was tested, and the prediction results were integrated into a decision framework for the business. In past research, cost estimation data sets generally consisted of <300 of projects from a range of a companies and even countries. This thesis case study only had access to the company's internal data but this contained over 2,000 past jobs Furthermore, the problem of cost estimation is applied to a consulting company in the construction industry, which differs from other studies in the construction industry that estimate actual construction costs. Consulting in the construction industry can be likened more to software development effort estimation although the nature of the work differs considerably.

ensemble tree methods and blended methods are applied which have been minimally tested on this problem. To date, Linear Regression, Neural Networks, Case Based Reasoning (CBR), and Support Vector Machines (SVM's) have been used to predict cost/effort even though ensemble tree methods can perform as well as Neural Networks and generally outperform Linear Regression (Caruana and Niculescu-Mizil 2006). Another benefit of ensemble trees is the output types available, such as partial dependency plots and variable importance plots that provide insight into the model's calculations. This contrasts Neural Networks and SVM's, which are 'black-box' predictors.

Finally, this study propels the predictive model one step further than other studies by analysing bottom-line profit improvements that could result from the research. final profit increases are evaluated. This kind of thought experiment is designed to present a clear case for industry when evaluating if and how to adopt cost estimation models.

3. Prediction Methods

Should provide sufficient detail to allow the work to be reproduced. Methods already published should be indicated by a reference: only relevant method modifications should be described.

Neural Netowrks and SVM were not purusued because black box A previously mentioned disadvantage is that the NN algorithm is a black box because the internal structure is too complex for interpretation. They also require a lot of training data relative to other methods.

Also, SVM's can be very slow to train and therefore not suitable for industry purposes (Auria and Moro 2008). Because it is so important to engage decision makers with a model that can explain its results, these two methods are not appropriate for the effort estimation problem.

3.1 Predictive methods. binary classification problems, where the objective is to predict the probability of an event occurring. Probability is a continuous response variable, however predictions from a linear regression would not be bound by 0 and 1.

3.1.1 Logistic Regression Because of the linear relationship, an equation can be fit to the log odds of the binary response variable against each of the explanatory variables using linear regression. Then, the linear equation describing the log odds can be transformed back to probability by taking the inverse log i.e. the exponential and rearranging (Macdonald 1975):

The result in a sigmoidal function bound by 0 and 1. The sigmoidal function originates from a linear fit of the log odds in the data. Assume linear relationship between covariates and response variable

A coefficient represents the change in the log odds of the response variable for each unit increase in an explanatory variable. Therefore, taking the exponential of the coefficient is the change in odds of the response variable. If the change in odds was 2, then if the explanatory variable increased by 1, the response variable event would be twice as likely to occur.

Logistic Regression is a good benchmark to compare other binary predictive models due to its simplicity and speed (Moore and McCabe 1989). single linear relationship means Not prone to overfitting

3.1.2 Random Forests single decision trees also suffer from low predictive accuracy and instability, so to combat these problems, ensemble tree methods were pioneered in the 1990's (Breiman 1996).

The three methods of combining hundreds or thousands of trees turn decision trees into high performing, stable predictors.

training multiple trees from the same data by sampling bootstrapped training sets with replacement. However, when creating each tree, a random subset of attributes (variables) is considered at each split. The reduced subset of attributes is resampled for each split in the tree. This allows dominant variables to be suppressed for a fraction of the splits, allowing the algorithm to explore signals in weaker variables (Breiman 2001).

Additionally, ensemble tree methods are faster than SVM's and Neural Networks but can perform just as well and even provide insights such as variable importance and variable relationships (Sealfon and Gymrek 2012).

3.1.3 Gradient Boosted Trees Lastly, the boosted decision tree approach applies gradient descent theory to a series of decision trees. The trees are limited to a certain depth to maintain simplicity, and each tree models the residuals (or errors) of the preceding tree. By modeling the errors, misclassified cases are weighted higher than correctly classified cases and influence the structure of the latest tree more (Hastie, Tibshirani, and Friedman 2009). This increased weighting on errors is why the algorithm is called boosting. The limited depth of each tree prevents overfitting at each stage and the combined result of up to thousands of trees is very powerful (Elith, Leathwick, and Hastie 2008).

Caruana and Niculescu-Mizil (2006) tested Boosted Trees, Random Forests, Neural Networks, SVM's, Logistic Regression and Naive Bayes on 11 binary classification problems and found that Boosted Trees performed best, followed by Random Forests. This demonstrates how ensemble tree methods are capable of competing with high-level machine learning algorithms.

Trees are not built on a probabilistic framework, and therefore their results cannot be provided in this framework (Louppe and Prettenhofer 2014).

3.1.4 Naive Bayes The Naive Bayes method works by making conditional independence assumptions about the explanatory variables in order to simplify probability calculations for the response variable (the response variable must be categorical).

These equations are simple to compute given the data, and the class with the highest probability can then be chosen (F. Provost and Fawcett 2013).

The advantages of this method are that the conditional independence assumption enables very fast calculations and predictions. The method can perform well in real world tasks because the assumption of independence does not significantly damage predictions (F. Provost and Fawcett 2013). Assumptions of independence genearlly not true (correlated variables). This is fine for ranking but the output probabilities are not accurate statistical probabilities (Caruana and Niculescu-Mizil 2006).

Naive Bayes method generally provides a good benchmark to compare against more complex models that should outperform it (Caruana and Niculescu-Mizil 2006).

3.1.5 Bayesian Networks A Bayesian Network is a graphical probabilistic model that illustrates the conditional dependencies between variables in a data set. The model is visually represented by a directed acyclic graph (DAG) and is capable of linking the conditional dependency between any variable to another variable (Heckerman 1998). The conditional relationships are Bayesian, where the probabilities in one node are conditional upon values in nodes directed towards it as well as preceding nodes

Bayesian Networks reduce computations required to find the probability of a unique combination given other explanatory variable values while still taking into account many conditional dependencies that Naive Bayes (Barber 2012)

Naive Bayes is the simplest form of a Bayesian Network (Zhang 2004)

Network relationships can be learned from the data, however this is not widely included as part of the suite of machine learning methods. The conditional variable dependencies are calculated from the data, which in turn can define the graph structure. Some conditional dependencies can be set before the structure is learned (Barber 2012).

Drawbacks of Bayesian Networks include the difficulty for them to process continuous variables.

Bayesian networks have found success in combining deterministic models with observed data as well as expert knowledge.(Kragt 2009)

3.2 Procedure. It includes first how the data was obtained, the lengthy cleaning process, Projects varied from total invoiced amounts of $500 to over $1,000,000 a project could have thousands of rows of relevant data that needed to be converted into a single row per project. 10 engineered variables describing timesheets: Percent of hours performed by 'professional' employees as opposed to 'technical' employees, Position of the employee that completed the most hours on each project 3 engineered from invoicing data: 4 from project data: number of projects completed with each client text analysis of project titles to develop project classifications

Discretisation was performed by generating a hierarchical dendrogram of each variable to visualise the clusters.

Limiting the predictive model to a concise set of meaningful variables reduces noise and improves predictions. Less variables and a simpler model is easier for stakeholders to understand (Weisberg 2005). followed by variable importance analysis, variable selection: varialbe importance aalyses done with linear regression, random forest, cforest to represent the range of algroithms that were triallled. 11 variables chosen:

trials of selected predictive algorithms. binary classification problem predicting profit or loss. Complex models should be measured against simple models that can be built at a fraction of the computational cost.

- Regression - baseline model
- Naive Bayes - baseline model
- Bayesian Network - grabage.

- Random Forest
- Gradient Boosted Trees

To compare the models, the area under the receiver operating characteristic (ROC) curve (AUC) statistic was used. An ROC graph visualises the curve from which an AUC score is calculated. In this problem a 'positive' is a loss making job. For binary classification, AUC is a more meaningful statistic than classification accuracy when the output is a probability that can be applied to the problem. The AUC indicates model performance across many thresholds while classification accuracy represents a single threshold.

a model must be first created using a training set. The model can then make predictions on the test set. this is used to build the ROC A model that is perfectly classified would have an AUC = 1. A model that predicts as well as random chance would have an AUC = 0.5. An AUC between 0.5 and 1 means the model is performing better than random chance.

In order to compare which models performed significantly better than others, an adequate sample size of results statistics is required. Multiple models could be made by using different data in the training vs. testing sets, each providing a resulting test statistic (RMSE or AUC). Initially, 20 models of each method were created in this fashion. Then a two-sample power calculation was run using the two sets of 20 results to determine the sample size to achieve a statistical power of 0.8 (100 to achieve power of 0.8 between boosted trees and naive bayes)

Missing Data Imputation. All methods except gradient Boosted Trees and Naive Bayes could not handle missing data. a complete data set allows for complete sets of predictions from each method, and these predictions could then be blended to further improve results. The MICE Random Forest method was chosen for imputation because it has been proven to work well with complex data sets (Shah et al. 2014). If similar predictive results were obtained using Boosted Trees imputed data and imputed data, the imputed data must be reasonable. The imputed data set was then trialed on the remaining methods and compared to unimputed trials.

Once the best methods were selected, they were blended in numerous ways, using both simple averaging techniques and sophisticated machine-learning algorithms. These were compared against individual models and the best constructs were selected. Six blending methods, ranging from simple to complex, were tested using predictions from the top performing individual models. These included simple averaging of the individual model results, building a Logistic Regression model using the individual model results only, a Boosted Tree model using individual model results only, feature weighted linear stacking (FWLS), Random Forests, and Boosted Trees. A simplified explanation of FWLS is as a Linear Regression where meta-features as well as model predictions from individual models are included as explanatory variables. Then, each meta-feature is interacted with each set of model results (Sill et al. 2009). Feature interaction is performed passively due to the nature of how trees are built. A split in a node determined by one variable is conditional upon the preceding split, which was based on another variable and so on. The simplest method averaged the probability output (a number between 0 and 1) of all three models. The next two simplest methods consisted of building a Logistic Regression and Boosted Trees model from the three prediction model outputs only. Feature Weighted Linear Stacking (FWLS), Random Forest and Boosted Trees were also tested as blending methods in a more complex scenario where the predictions from each model became additional variables to the original explanatory variables (called meta-features in this context). The model predictions were interacted with each original variable so that models that performed more strongly under certain meta-feature states could be weighted as such.

Again, 100 models were built to achieve a power of 0.8 for the variation in AUC test statistics across different divisions of training/testing data. A maximum of five blended models were created from a complete set of test results (built from the individual models), then a new set of test results were created for the next five blended models.

Finally the impact of the algorithm on the overall profits of the case study company was analysed via decision-making scenarios. This analysis differed from accuracy in predicting profitability, so several blended models as well as individual models were carried forward for this analysis. This was done using a profit curve - a chart that plots the change in profit the company earns on the y-axis vs. the probability threshold on the x-axis. A simple approach was taken for this analysis, where projects with a probability to be a loss making job greater than the threshold were rejected entirely. Therefore all profits and losses from jobs above the threshold were discounted.

An equation defining the change in profit as a percentage of the original profit using threshold rejection is shown below:

$$\Delta\ Overall\ Profit\ (\%) = \frac{\sum_{p=1}^{N} I(Pr_p < threshold) \cdot Profit_p}{\sum_{p=1}^{N} Profit_p} \cdot 100 \tag{1}$$

Where

$I(\cdot)$ = the indicator function

N = the number of individual projects that are being included in the analysis

Pr_p = probability output from the algorithm for project p. Values are betweeen 0 and 1 where 1 is loss making)

$threshold$ = a chosen value between 0 and 1. $\Delta\ Overall\ Profit$ is calculated for several $threshold$ values which defines the profit curve

$Profit_p$ = profit for individual project p

If the threshold was zero, all jobs were rejected and the profit would be \$0. If the threshold was 1.0, all jobs were accepted and the profit would be the same as the profit the company actually experienced since the data is a sample of historic projects. The aim was to find the optimal threshold point where saying ‘no’ to a job above that level would result in higher profits, because jobs that were likely to make a loss were being rejected. This chart will clarify what percentage of profit increase the company could expect by integrating the algorithm into decision-making.

the curve varies with different divisions of the data. Therefore, in order to understand the uncertainty around the profit curve and to determine which blended or individual model performed statistically better than others, a large sample size of curves was required. 100 to achieve a power of 0.8. which was achieved by repeating 5-fold training/testing splits. A 95% confidence interval could also be determined around the highest point on each curve. The final expected increase in profit and the percentage of projects to be rejected presented a clear scenario that the case study business managers could assess in terms of their business strategy.

4. Prediction Results and Discussion

Present results clearly and concisely discussion should explore the signficance of the results of the work, not repeat them.

4.1 Individual Models. The predictive formula was re-structured so that the 11 explanatory variables predicted a new response variable: profitable/unprofitable projects. The five prediction methods, were initially built without imputing data, by using the maximum amount of data possible depending on the method. Boosted Trees and Naive Bayes are able to process data with missing values, so all data could be input (Ridgeway 2015; Meyer et al. 2014). 100 models of each method were built. The violin plot below summarises the AUC values produced by each method and the ‘violins’ are coloured according to whether the distributions significantly vary to Boosted Trees using a critical value of 0.05.

The AUC performance of Logistic Regression and the Random Forest algorithms cannot be statistically differentiated from Boosted Trees.

Next, data imputation methods were trialed which would make the complete data set available to Logistic Regression and Random Forests. Again, 100 models were required to achieve a power of 0.8 with respect to Boosted Trees

Boosted Trees, Logistic Regression, and Random Forest all performed significantly better than the baseline algorithm, Naive Bayes, however none outperformed Boosted Trees.

Testing Boosted Trees, Random Forests, Naive Bayes, Logistic Regression and Bayesian Networks on the binary classification problem showed that Boosted Trees, Logistic Regression and Random Forests performed best (according to AUC). Naive Bayes and Logistic Regression were included as baseline models and it was expected that the more complex models would outperform these. Therefore it was surprising that results from 100 Logistic Regression models were not statistically significantly lower than Boosted Trees.

A possible explanation for this, according to the literature, is that there was not enough data for the ensemble trees to learn the complex decision rules at which they excel. Trees tend to overfit the patterns in a smaller training set. Logistic Regression on the other hand is capable of only one decision boundary (which does not have to be parallel to the variable axes) and is not prone to overfitting (Perlich, Provost, and Simonoff 2003). This may explain Logistic Regression's comparatively high performance on the case study's small but complex data set.

Boosted trees, Random Forests and Logistic Regression all had mean AUC values between 0.755 to 0.764. In summary, the individual binary classification models performed well above random chance (AUC = 0.5). Whether the model is worth implementing in the work place is dependent on the extent to which the algorithm would improve 'bottom line' profits for the business and if the model can affect decisions in practice.

- NB and BN were bad - one sentence

4.2 Blended Models. Blended models combine the predictions from each high-performing individual model to create averaged or 'blended' predictions. In this case, the Random Forest, Boosted Trees, and Logistic Regression models were chosen. Through variable importance studies, 6 variables were included as metafeatures. All six methods were compared against the original Logistic Regression model's AUC distribution (as it was not statistically different from the ensemble trees)

The above plot shows that 4 methods had a higher mean AUC than the original Logistic Regression model:

- the simple blended Logistic Regression
- the simple average
- and the two Boosted Tree models.

The simple averaged method and simple logistic regression could achieve a statistical power of 0.8 with 150 samples (models), whereas the boosted trees required thousands. The additional 50 models were run for the simple logistic regression and simple average to reveal similar distributions that were indeed statistically significantly higher than the individual logistic regression AUC distribution. P-values from two sample t-test were 0.0019 and 0.0021 respectively.

Blended models improved the mean AUC from 0.759 to 0.77, which is an increase of 1.4%. It was expected that model blending would improve predictions because it combined the strengths of individual models with different theoretical foundations.

The trials of different blending methods demonstrated that again, the simplest methods worked best with the case study data. In this incident, averaging the results of the best three single models (Logistic Regression, Random Forests, and Boosted Trees) or taking a Logistic Regression of the three models outperformed more complex Boosted Trees, Random Forests, and Logistic Regression blends that facilitated interaction of the individual model results and original variables (meta-features). Simple blended models could perform better than complex methods if the data is not big enough for the complex models to learn the more intricate patterns at which they excel. This was observed in the single model methods. their success in the 2009 Netflix competition generated some publications. The Netflix data set comprised of almost 3,000,000 observations, so the size of the data could have enabled the success of complex blending methods such as FWLS (Sill et al. 2009)

5. Business Impact

This chapter first presents the full range of method results in terms of improvements to the case study's bottom line. A business decision making scenario was created so that profit curves based on the decision rule could be built and analysed. From the profit curves, optimal probability thresholds could be derived for each method (individual and blended methods). The predictive models output a 'probability' between 0 and 1 that each project will be a loss making job (where probability = 1 indicates a loss making job). The question the arises, at what probability would a decision-maker round the probability up to 1 or down to 0? And what business decision would then be made? To find the threshold point for rounding, an experimental business-scenario was tested. At a given threshold, all projects with probability outputs above the threshold were considered too risky, and were rejected. All profits and losses from these projects were forfeited, while the profits from the remaining jobs (below the threshold) were summed to give a revised total profit. This profit calculation was made on a range of thresholds between 0 and 1 at 0.05 increments. The total profits were plotted for each threshold and joined to make a profit curve

The plot below illustrates the distribution of profit curves for each method, where the solid lines join the mean values at each threshold point. The grey ribbon illustrates the 95% confidence interval for the profit improvement ratio at each threshold point for the 100 models.

Two blended models clearly outperformed the individual models as shown in the above plot with higher profit ratios as well as tighter confidence intervals. The simple Logistic Regression blend performed best with the highest mean profit ratio of 109% with a standard deviation of 1.28%. %. This means that for the simple blended Logistic Regression model, if all jobs above the probability threshold 0.6 were rejected, the profits would increase on average by 9% in comparison to historical profits.

Profit curves from the simple average blend and Logistic Regression blended models outperformed the complex blended models, which follows logically from their significantly higher AUC distributions. The simple average blended model produced a profit curve with an almost identical profit improvement (and standard deviation) to the Logistic Regression blend. In the data, projects assigned a probability higher than 0.6 represent only 4.3% of all projects.

The shaded grey 95% confidence intervals in the Profit Curve plot shows that with the 100 sample models that were analysed, none of the lower bounds for the original methods' or complex blending methods were above 100% on the y-axis. That is, it cannot be said with 95% confidence that the original methods would produce a mean profit higher than historical profits in the given decision framework. Clearly this level of certainty is not beneficial for the case-study business to adopt.

It is not clear why simple linear models and the averaging model outperformed ensemble tree blending methods. As previously stated, the ensemble trees may not have received enough data to adequately learn the complex series of rules they develop. The only differences between FWLS and the simple Logistic Regression (that performed best) were two additional explanatory variables and four interaction terms. The number of variables was not high for the amount of data since according to Peduzzi et al. (1996), 10 events per explanatory variable or more avoids the risk of biased estimation of variable coefficients in Logistic Regression. The data contained 315 events per explanatory variable and did not pose that risk. Nevertheless, the additional variables must have added misleading noise to the model.

To conclude, because of the promising AUC results, it was logical that the models translated into financial benefits for the company. The 9% improvement in profits produced by the simple blended Logistic Regression is reasonable, and may be high enough to trigger further cost-benefit analyses and the development of a more comprehensive framework describing different decision scenarios.

- decision scenario
- profit curves
- prototype decision support interface

6. Conclusions and Future Work

3/4 page

The general aim was to use statistical techniques to predict the profitability of projects for a case study consulting business using their internal CRM data. This was rigorously completed by approaching the prediction of 'profitability' as a binary classification problem predicting either profit or loss. Several statistical and machine learning approaches were applied including Naive Bayes, Bayesian Networks, Linear Regression, Random Forests, gradient Boosted Trees as well as multiple methods of blending the individual models' output.

AUC = 0.76). This was achieved by 3 individual methods while various techniques that blended the individual methods improved results further (AUC = 0.77).

predictive models could be shown to improve the overall profitability (bottom line) of the case study business. A range of probability threshold values were trialed for each method using a decision framework where projects scored below the threshold were accepted, while projects above the threshold were rejected. If a project was rejected, the profits and losses were forfeited, and the remaining accepted profits and losses were summed.

Final results showed the simple Logistic Regression blend of the individual Logistic Regression, Random Forest and Boosted Tree models improved profits the most. The 95% confidence interval for these improvements was between 6.5% and 11.5% using a probability threshold of 0.6 (approximately 4.3% of projects).

These results contribute significantly to the research in cost estimation in three ways: the applied methods, the decision framework, and the appeal to user trust. Ensemble tree methods and blending had been applied minimally to cost estimation previously, even though ensemble trees provide insight into model structure while predicting at a similar level to Neural Networks. Next, previous studies have verified predictive accuracy but stopped short of how the algorithm would affect decisions and what the measured benefits would be. This study presented a clear framework for how a business could improve profits by applying the algorithm.

Further work is required to test user confidence in the output. Another topic identified for future research was to test how well managers estimate some numeric project input variables. In particular, time span, team size, total invoiced amount, and percentage of hours completed by professionals should be tested as these variables were calculated post project completion. Time span and total invoiced amount were discretised into wide categories, which should be easier for a manager to choose between.

Overall, this work has successfully built a mathematical blend of Logistic Regression, Random Forests, and Boosted Tree models, from a consulting company's internal project data. This blended model can predict whether a project will be profitable or not and in a reasonable decision framework, can guide managers in rejecting financially risky projects and improving profitability of the business.

Acknowledgments

This research was supported by a scholarship/ ACEMS?

References

Installation. If the document class *elsarticle* is not available on your computer, you can download and install the system package *texlive-publishers* (Linux) or install the LaTeX package *elsarticle* using the package manager of your TeX installation, which is typically TeX Live or MikTeX.

The author names and affiliations could be formatted in two ways:

(1) Group the authors per affiliation.
(2) Use footnotes to indicate the affiliations.

Bullet points.

- document style
- baselineskip

- front matter
- keywords and MSC codes

Akintoye, Akintola, and Eamon Fitzgerald. 2000. “A Survey of Current Cost Estimating Practices in the UK.” Journal Article. *Construction Management and Economics* 18 (2): 161–72. doi:10.1080/014461900370799.

Attalla, Mohamed, and Tarek Hegazy. 2003. “Predicting Cost Deviation in Reconstruction Projects: Artificial Neural Networks Versus Regression.” Journal Article. *Journal of Construction Engineering and Management* 129 (4): 405–11.

Auria, Laura, and Rouslan A Moro. 2008. “Support Vector Machines (SVM) as a Technique for Solvency Analysis.” Journal Article.

Badenfelt, Ulrika. 2011. “Fixing the Contract After the Contract Is Fixed: A Study of Incomplete Contracts in IT and Construction Projects.” Journal Article. *International Journal of Project Management* 29 (5): 568–76. doi:http://dx.doi.org/10.1016/j.ijproman.2010.04.003.

Barber, David. 2012. *Bayesian Reasoning and Machine Learning*. Book. Cambridge University Press.

Breiman, Leo. 1996. “Bagging Predictors.” Journal Article. *Machine Learning* 24 (2): 123–40.

———. 2001. “Random Forests.” Journal Article. *Machine Learning* 45 (1): 5–32. doi:10.1023/A:1010933404324.

Brown, Iain, and Christophe Mues. 2012. “An Experimental Comparison of Classification Algorithms for Imbalanced Credit Scoring Data Sets.” Journal Article. *Expert Systems with Applications* 39 (3): 3446–53.

Caruana, Rich, and Alexandru Niculescu-Mizil. 2006. “An Empirical Comparison of Supervised Learning Algorithms.” Conference Proceedings. In, 148:161–68. ACM. doi:10.1145/1143844.1143865.

Chan, Swee Lean, and Moonseo Park. 2005. “Project Cost Estimation Using Principal Component Regression.” Journal Article. *Construction Management and Economics* 23 (3): 295–304.

Elfaki, Abdelrahman Osman, Saleh Alatawi, and Eyad Abushandi. 2014. “Using Intelligent Techniques in Construction Project Cost Estimation: 10-Year Survey.” Journal Article. *Advances in Civil Engineering* 2014.

Elith, Jane, John R Leathwick, and Trevor Hastie. 2008. “A Working Guide to Boosted Regression Trees.” Journal Article. *Journal of Animal Ecology* 77 (4): 802–13.

Finnie, Gavin R, Gerhard E Wittig, and Jean-Marc Desharnais. 1997. “A Comparison of Software Effort Estimation Techniques: Using Function Points with Neural Networks, Case-Based Reasoning and Regression Models.” Journal Article. *Journal of Systems and Software* 39 (3): 281–89.

Flyvbjerg, Bent. 2007. “Cost Overruns and Demand Shortfalls in Urban Rail and Other Infrastructure.” Journal Article. *Transportation Planning and Technology* 30 (1): 9–30.

———. 2011. “Over Budget, over Time, over and over Again: Managing Major Projects.” Journal Article.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Book. Springer New York. http://qut.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwzV1bS8MwFA7iQBwD3dR6mRAQfNk6trb0Ivi0TQY6EJwgQxi9JLKHddpVf7_nJE269cFnH9NCm_Y7JF--cyPEtnp9s7Im8JCzAWxVjFluEnMWBF4MyyfzE4-HVsR29Tfd9rO89s-BrzBiPSwRH8tocBGfgTxSlG
lQg53OGzhsR9IHSKJ6YsfDxcttlP8vYzzrrTEJ4vjYNGaUN8EcoeYfpUofnMCyKACQZ-O6qMEelNAQVpUEpjR1VUGrnKIpucN_
BcvnlzqK86ZUNR4cBwrbpes4tFjlf4Yfds9R8fYH9dAA8FGnufK71MnR_ArkqHO7iTW5RQEm_uU4aMEPxsA9subLFFGbHpMY
iUBUYW8r_YZ6QRYppDmot0yMQgNQ6myQzkCQZ8l0EO3oKnkT95HMphUw17G5Gz1_vKDaAlwrJNt-edE2pb3PEiL-aRkzh-AKfFqO8nPse2BbFt9S9I-69JXf59-4oclpbTJvt59s2uRQ7tL6t_R8w.

Heckerman, David. 1998. *A Tutorial on Learning with Bayesian Networks*. Book. Springer.

Kim, Gwang-Hee, Sung-Hoon An, and Kyung-In Kang. 2004. “Comparison of Construction Cost Estimating Models Based on Regression Analysis, Neural Networks, and Case-Based Reasoning.” Journal Article. *Building* *and Environment* 39 (10): 1235–42. doi:10.1016/j.buildenv.2004.02.013.

Kragt, Marit E. 2009. *A Beginners Guide to Bayesian Network Modelling for Integrated Catchment Management.* Book. Landscape Logic.

Kumar, P Ravi, and Vadlamani Ravi. 2007. “Bankruptcy Prediction in Banks and Firms via Statistical and Intelligent Techniques–A Review.” Journal Article. *European Journal of Operational Research* 180 (1): 1–28.

Louppe, Gilles, and Peter Prettenhofer. 2014. “Gradient Boosted Regression Trees.” Online Multimedia. PyData.

Lovallo, Dan, and Daniel Kahneman. 2003. “Delusions of Success: How Optimism Undermines Executives’ Decisions.” Generic. HARVARD BUSINESS SCHOOL PUBLISHING CORPORATION. http://qut.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMw3V3Pb9MwFLYQQxMSEgxYKAPJBwQHlCqO61-TdoC2aNImcdgmcYscxwGktoMlQfvzeXYSJy1M4swxjZuq__uz3vs95PxCi6TSJd2yCAU-T53pmjbIgoblMCqGYyEtpEpoWs-3zt9Bqb__jsfwB-YVdN1Ye3XTS-IaLT__a553GcwEGvXFsO3O1q7kPf3y1trGmf0fDzFouu5U41Za99AKETJb79ROL__-pbtXOIthsZ3pbxvbna-2iexbJww0RKMGq9l5srHVlGS0OsTIBDI-cqakTczcrnO9439CVCCIZQ6ml751Vc__XxXdTn9hNfHUBDhZ4q1cFH5d3utk__feqIVDxwOT5N9Vd24ZnE5RP0uC__pjT-02B2ge3bzFO33k__sMzQOE-LrEHYTHGADEPYB4ABAPAL7DAb7n6OrT8nJ-GnfNLuKvBChkzFTJcgnkVadGg1XkVllKiXYEDHYLYxKUK3imPCfGklIwA8
izpfyNOzeXt50F9OK5__hN__1ZRzC7fh__EfCpeICxEIg2fcebqtoHtzlOic26UYobKWZFMUOQmPXObor7RJiMJT0BSJGqCDnscsn
cOABY3jDfSdkYIhgzEueCSL__MmzeFcl3xSHql3f-6BF6OGyaV-h-fdPY1z5t-DdLboxJ.

Macdonald, P. 1975. “The Logit Transformation: With Special Reference to Its Uses in Bio-Assay.” Journal Article. *Journal of the Operational Research Society* 25 (1): 201–2.

Matson, Jack E, and Joseph M Mellichamp. 1993. “An Object-oriented Tool for Function Point Analysis.” Journal Article. *Expert Systems* 10 (1): 3–14.

Mendes, Emilia, and Barbara Kitchenham. 2004. “Further Comparison of Cross-Company and Within-Company Effort Estimation Models for Web Applications.” Conference Proceedings. In *Software Metrics, 2004. Proceedings. 10th International Symposium on*, 348–57. IEEE.

Meyer, David, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. 2014. *E1071: Misc Functions of the Department of Statistics (E1071), TU Wien.* https://CRAN.R-project.org/package=e1071.

Moløkken, Kjetil, and Magne Jørgensen. 2003. “A Review of Software Surveys on Software Effort Estimation.” Conference Proceedings. In *Empirical Software Engineering, 2003. ISESE 2003. Proceedings. 2003 International Symposium on*, 223–30. IEEE.

Moore, David S, and George P McCabe. 1989. *Introduction to the Practice of Statistics.* Book. WH Freeman/Times Books/Henry Holt & Co.

Moores, TT, and JS Edwards. 1992. “Could Large UK Corporations and Computing Companies Use Software Cost Estimating Tools?–A Survey.” Journal Article. *European Journal of Information Systems* 1 (5): 311–20.

Pai, Dinesh R., Kevin S. McFall, and Girish H. Subramanian. 2013. “Software Effort Estimation Using a Neural Network Ensemble.” Journal Article. *Journal of Computer Information Systems* 53 (4): 49–58. http://qut.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwnV1Nb9QwEB0VekFClPLVAJV8gGPAH4kbc4GlyraIbhftpurRysTe06os3VTqz8fjTl
Ej-WC8rFHnUE4x__Fmlv__e1WUh7I1fKfFiham-EO9Nuf-tjX__FZywW02x3t__xFN40nmlbLAaRruw5S-ewU6v-MA6AHgOn6bjYXU-mJQs-LfjScXKafVtFMNcjAQ8jtiAnZZnk8FJOFTn48l3Vp5Oy9HXk__IFnA3L6vA47dQX0gWlFaWNwjwjO
FxTbv0L9qYzef2gKkw-eJZ1BnFzGcKQ__Ua3SxUyRELTGC__byZbI8V6mnZpbxopgVfr624-t2G-QynmSooE3q-6zy5WLB1W2qW03BYBoSTdrbltr9sE9jbKkaChKLgyCbz7s-PXBWjaGJ6guInGPQFxl2KHHcU6UQu0r____73m__gkYwCG7QB-C08bC-v__H5MOv0NwrD2yw.

Peduzzi, Peter, John Concato, Elizabeth Kemper, Theodore R Holford, and Alvan R Feinstein. 1996. “A Simulation Study of the Number of Events Per Variable in Logistic Regression Analysis.” Journal Article. *Journal of Clinical Epidemiology* 49 (12): 1373–9.

Perlich, Claudia, Foster Provost, and Jeffrey S Simonoff. 2003. “Tree Induction Vs. Logistic Regression: A Learning-Curve Analysis.” Journal Article. *The Journal of Machine Learning Research* 4: 211–55.

Pinto, Jeffrey K, and Dennis P Slevin. 1988. “Critical Success Factors Across the Project Life Cycle.” Conference

Proceedings. In. Project Management Institute.

Provost, F., and T. Fawcett. 2013. *Data Science for Business*. Book. O'Reilly. https://books.google.com.au/books?id=_1b4nAEACAAJ.

Ridgeway, Greg. 2015. *Gbm: Generalized Boosted Regression Models*. https://CRAN.R-project.org/package=gbm.

Saradhi, V Vijaya, and Girish Keshav Palshikar. 2011. "Employee Churn Prediction." Journal Article. *Expert Systems with Applications* 38 (3): 1999–2006.

Sealfon, Rachel, and Melissa Gymrek. 2012. "Recitation 6: Random Forests and Affinity Propagation." Online Multimedia. MIT University. https://stellar.mit.edu/S/course/6/fa12/6.047/courseMaterial/topics/topic4/lectureNotes/recitation6/recitation6.pdf.

Shah, Anoop D, Jonathan W Bartlett, James Carpenter, Owen Nicholas, and Harry Hemingway. 2014. "Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study." Journal Article. *American Journal of Epidemiology*, kwt312.

Shane, Jennifer S, Keith R Molenaar, Stuart Anderson, and Cliff Schexnayder. 2009. "Construction Project Cost Escalation Factors." Journal Article. *Journal of Management in Engineering* 25 (4): 221–29.

Shepperd, Martin, Chris Schofield, and Barbara Kitchenham. 1996. "Effort Estimation Using Analogy." Conference Proceedings. In *Proceedings of the 18th International Conference on Software Engineering*, 170–78. IEEE Computer Society.

Shin, Y. 2015. "Application of Boosting Regression Trees to Preliminary Cost Estimation in Building Construction Projects." Journal Article. *COMPUTATIONAL INTELLIGENCE AND NEUROSCIENCE* 2015: 149702. doi:10.1155/2015/149702.

Sill, Joseph, Gábor Takács, Lester Mackey, and David Lin. 2009. "Feature-Weighted Linear Stacking." Journal Article. *ArXiv Preprint ArXiv:0911.0460*.

Trost, Steven M, and Garold D Oberlender. 2003. "Predicting Accuracy of Early Cost Estimates Using Factor Analysis and Multivariate Regression." Journal Article. *Journal of Construction Engineering and Management* 129 (2): 198–204.

Weisberg, Sanford. 2005. *Applied Linear Regression*. Book. Vol. 528. John Wiley & Sons.

Zhang, Harry. 2004. "The Optimality of Naive Bayes." Journal Article. *AA* 1 (2): 3.

List of Figures

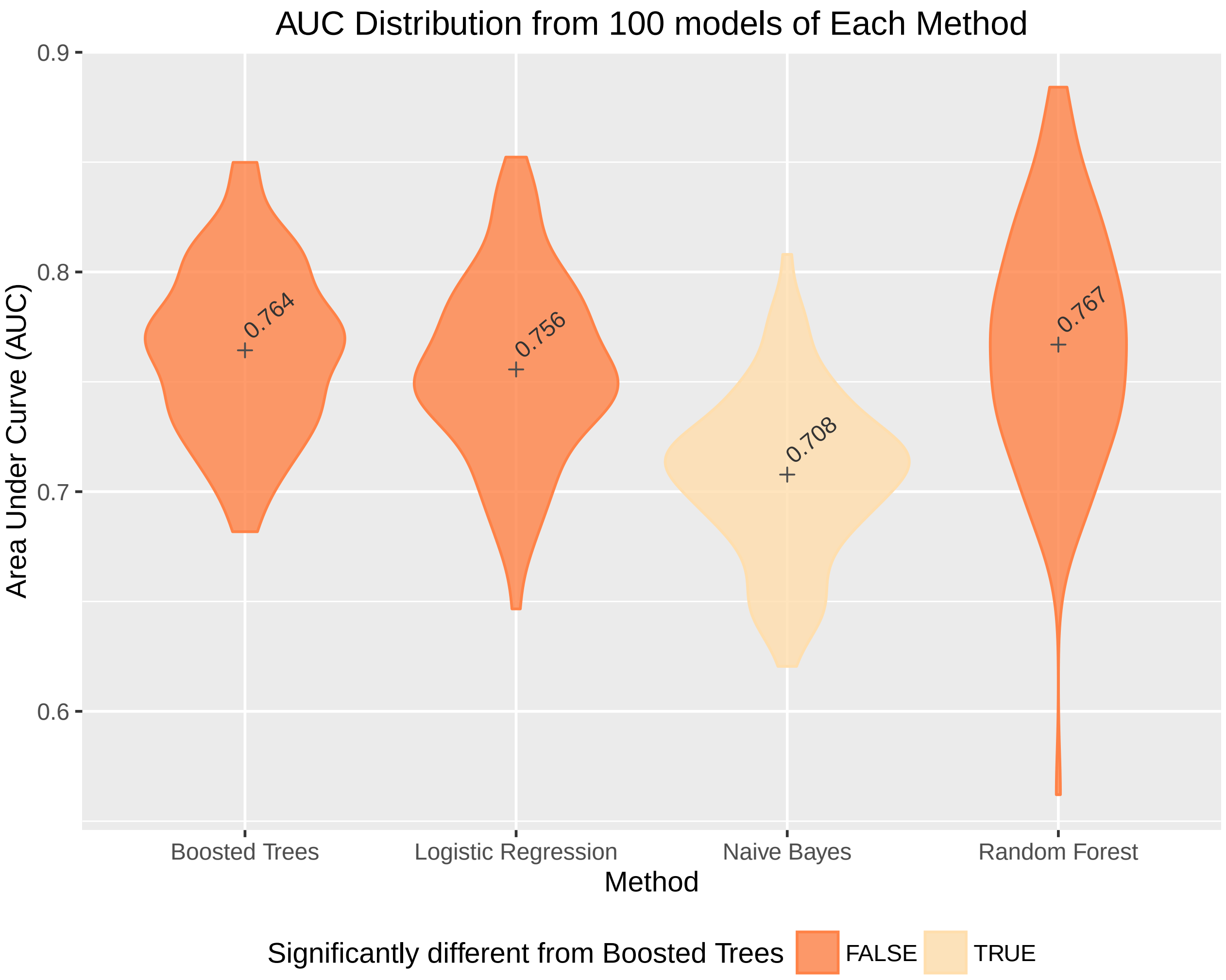

Figure 1: Violin plot vertically illustrating the distribution of AUC values from each of the methods when predicting profit/loss. Subsets of the data were used for Logistic Regression and Random Forests in order to provide datasets without missing values.

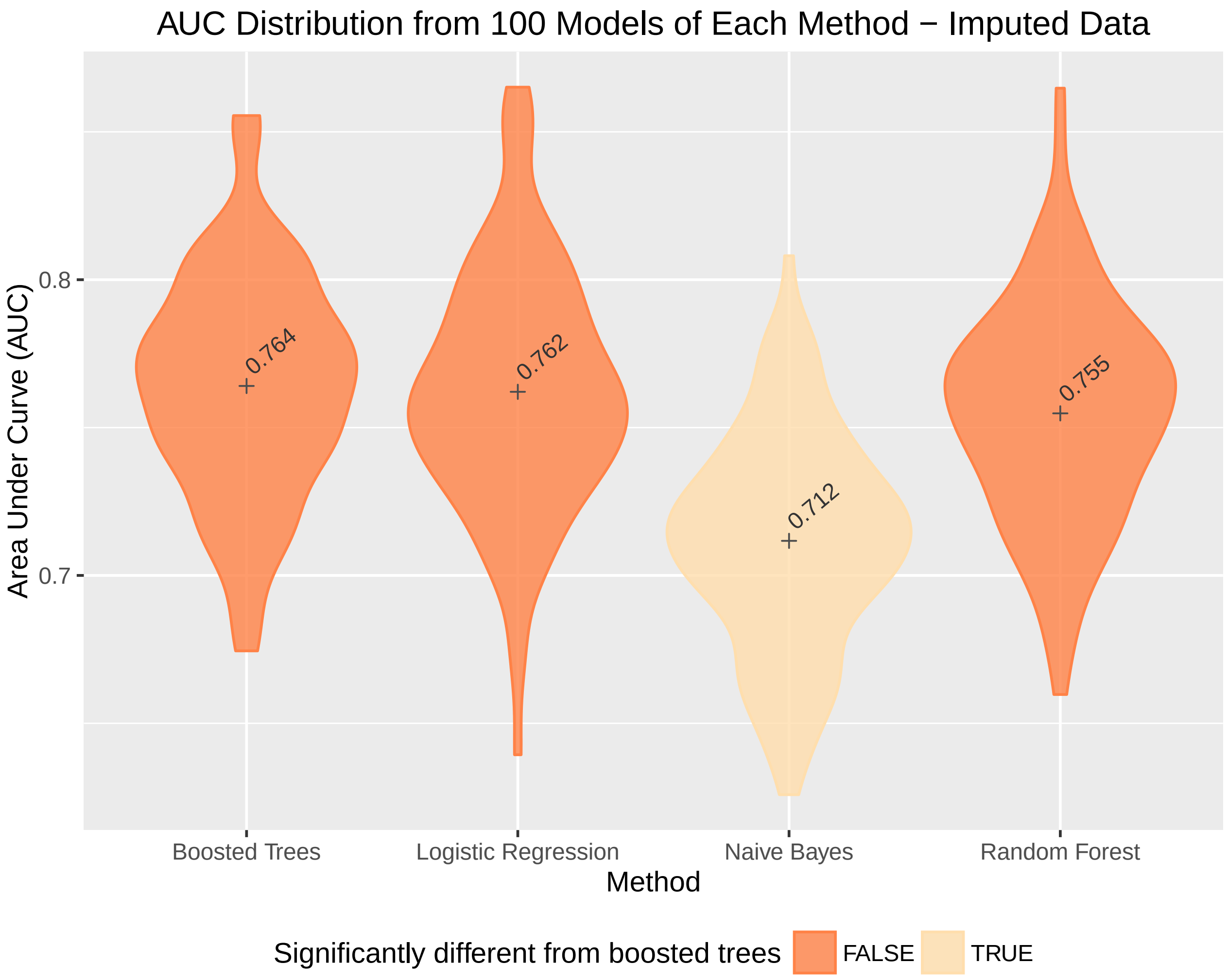

Figure 2: Violin plot vertically illustrating the distribution of AUC values from each of the methods when predicting profit/loss. Each method was fed the same imputed full dataset.

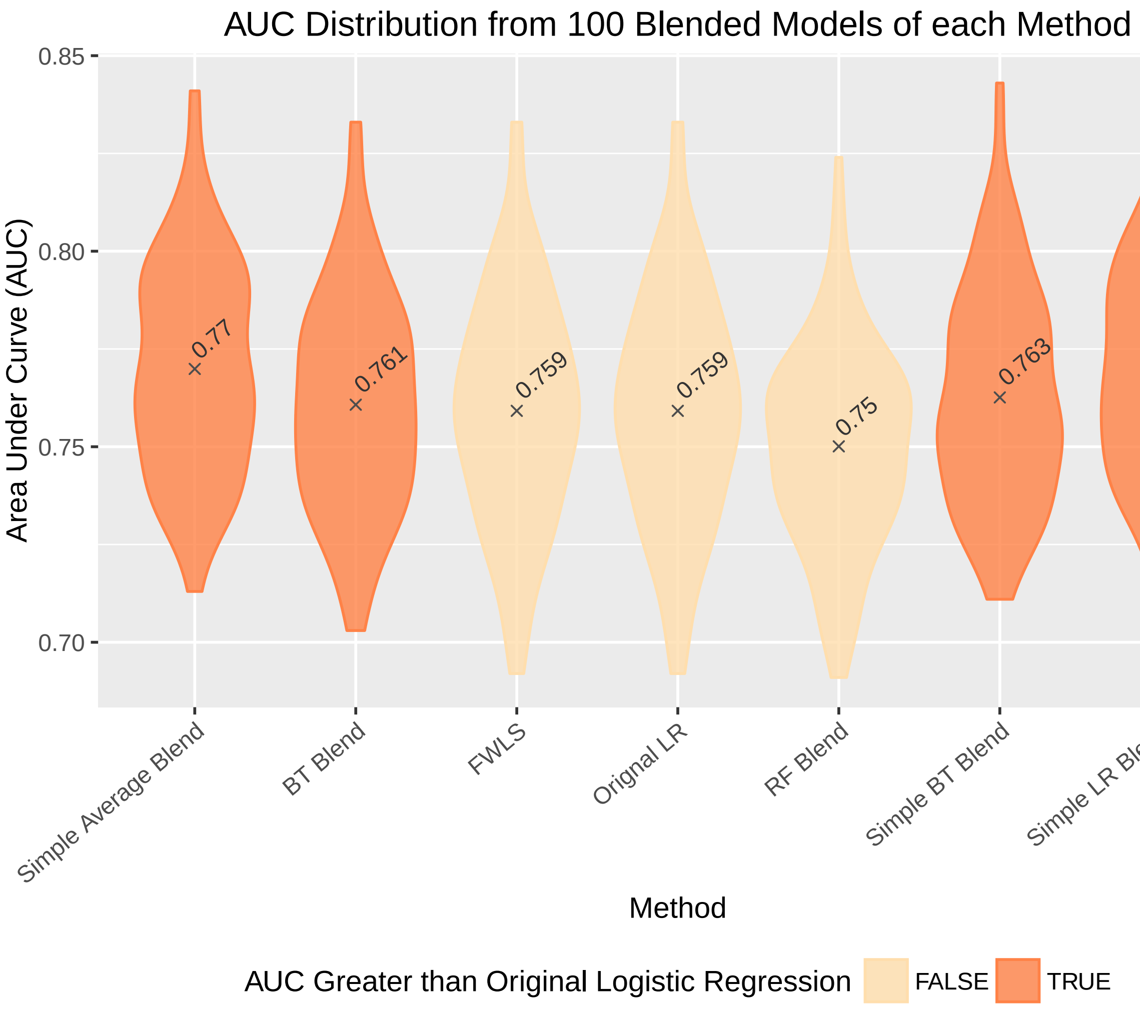

Figure 3: Violin plot vertically illustrating the distribution of AUC values from each of the blending methods when predicting profit/loss. 100 models were built for each method.

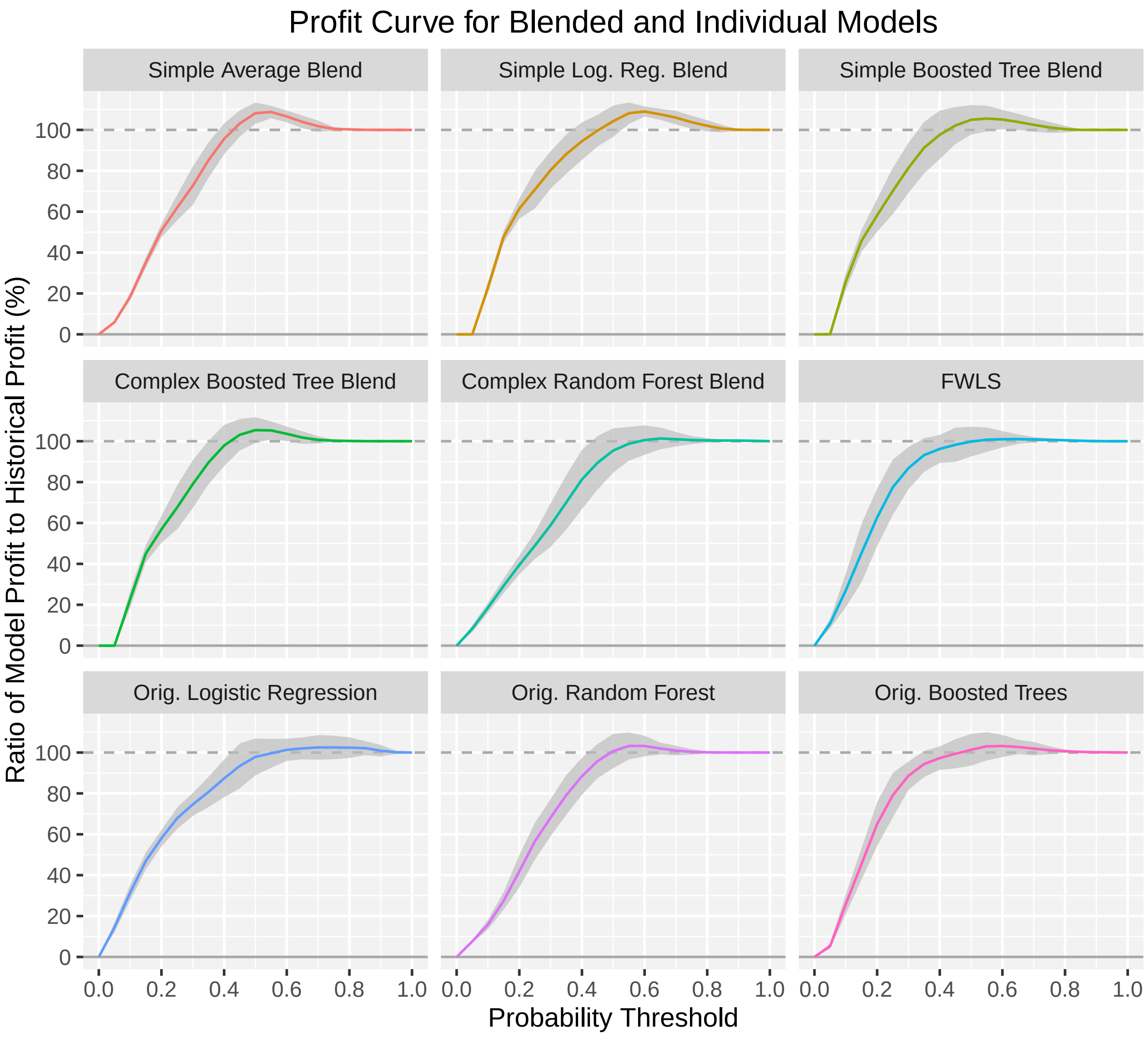

Figure 4: Profit curves summarising results from 100 models of 9 methods: 3 simple blends, 3 complex blends, and the original 3 best methods.