draft 1

Amy Cook

Tuesday, July 07, 2015

Originality Statement

'I hereby declare that this submission .... look up QUT'

Signed ................. Date....................

# Abstract

# Contents

# List of Figures

# List of Tables

# Glossary

B2B - Business to Business CRM - Customer Relationship Management. Refers to a common type of software used by businesses to record client and project details as well as employee timesheet records. Generally accessable to each employee in a company.

# Acknowledgements

# Publications

# Introduction - 2 pages

Predicting how long it will take a person to do something is a notoriously tricky task. People make mental calculations of these sort every day, guessing how long it will take to drive to meet someone on time, how many tasks you can check off in a day, or how long it will take to cook a dish. These mental calculations can be particularly inaccurate if the individual has never performed the task before. These errors are trivial in everyday activities, but they can have more meaning for businesses that sell their time as their chief product. For example, consulting businesses give expert advice to other professionals, in exchange for a fee for the amount of time the constultant spent on the problem. The delivery of 'expert advice' is generally in the form of a document (which could be anything from a report to hundreds of drawings) and can range from an hours advice to thousands of hours work on a substantial project. Before a project commences, the client and consultant must agree on a fee for the expert advice, or a fee structure. The nature of how a fee is structured can be as creative as the engaged parties wish, however a couple of simple examples are as follows: the client agrees to pay for the consultants time by the hour until the task is completed, or the consultant provides a fixed fee to complete the project in full, regardless of their hours spent. The nature of the fee structure brings an element of financial risk to one party which must be mitigated.

This thesis focuses on the risk a consulting company undertakes when setting a fixed fee for a project before their role has commenced. Clients commonly ask for competitive fixed quotes from several consultants before settling with one or negotiating further. The way a consulting manager decides on their fixed fee varies greatly from industry to industry and company to company. Typically, consulting manager has extensive experience in the type of project he is quoting and can use a combination of intuition, ratios of the entire project cost, and looking up a couple of past similar projects. Such a method often results in a wide range of profitabilities for projects depending on the ability of the manager to predict the amount of time the project will demand from the consulting team. This thesis examines a case study consulting company that has recorded internal historical project performance data since 2002. The study tests whether a statistical predictive model based on past project data can predict the profitability of a project based on characteristics available at the beginning of a project. This thesis asks further whether the prediction can impact a consulting manager's determination of a fixed fee.

# Problem Description

In this chapter the research motivations are explained along with the goals of the project and the specific research hypothesis. Finally the contributions of this body of work are outlined along with the structure of the thesis.

## Research Motivation - 4.5 pages

**Problem Description:**

Project decision makers have struggled with forecasting project costs accurately for decades, across many industries. A study on large-scale infrastructure projects over the past seventy years revealed that cost forecasts consistently underestimated the cost of rail projects by 44.7%, the cost of bridge and tunnel projects by 33.8% and the cost of road projects by 20.4% (Flyvbjerg, 2006). 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 on average (Flyvbjerg and Budzier, 2011). These statistics cast a concerning glimpse at the financial story playing out behind the scenes of these projects which can cause job losses, ruin private businesses, and churn through government balancesheets. Complex long term projects are necessary to produce railways, houses, websites, and movies. This research focuses on the financial risk that is transferred to consulting companies who are enlisted to complete portions of these complex projects. The risk is transferred by requiring a fixed flat fee from the consultant, meaning they commit to complete their role for a nominated price. In a similar way to infrastructure and IT projects, consultants are susceptible to underestimating the amount of their time that will be spent in their role, and must wear the cost of extra employee hours spent. For privately owned businesses, mitigating this risk would increase the chances of the business surviving as well as growing and succeeding.

The ramifications of unprofitable projects in a private business are multifold. Employee morale deteoriates and employees may view themselves as incompetent if they are responsible for the project's delivery. Employees also experience stress attempting to complete the project within a disappearing budget and may produce lower quality work in exchange for speed. This is unfortunate if in reality, the project fee was under estimated. Unprofitable projects also limit a business' opportunity to invest money into marketing or staff training, which would improve a business' standing. The financial risk taken on fixed flat fee projects may also discourage business' from taking different risks with better potential pay-offs, such as expanding the business into a new area. Logically, predicting fixed fee project costs more accurately could positively affect a business' growth opportunities and staff morale.

**Why its Not Easily Solvable:**

The challenge for consultants to reduce the risk of a project going over budget, in terms of hours spent, is not easily solved. Complex projects are always different from each other, and similar work may take significantly different amounts of time. For example, an ad filmed in a remote location for chocolate will cost significantly more than an ad filmed on an exisitng set, even though the product structure is almost identical. Therefore, a fixed product price, which you could apply to products such as appliances, is inappropriate for film projects. One option is charging for consultants' time and the hiring of materials by the hour; a zero risk fee structure. However in many industries this is not competitive enough. Other consultancies may be willing to offer a fixed flat fee for the same work, reducing financial risk for the client, and luring the client to them. Another fee strategy may be to track worker hours, and stop work once the consulting fee has been depleted. Then, negotiate further fees, or variations, with the client and persuade them to accept. This also satisfactorily reduces consultant financial risk, and in some cases, depending on the nature of the work or uncontrollable changes, this method can be effective. However, if the project has been following a predictable process but the hours are nevertheless inflating beyond expectation (a common situation), consultants may prefer to wear losses or endure very marginable profits to preserve the relationship with their client. Consultant's may believe it is in their best interest to maintain a reliable, trustworthy reputation in their industry over creating financial friction. Clearly, there is no straightforward solution to reducing the financial risk taken on by consultant's who offer fixed fees in complex projects. In some indusries, fixed flat fee structures are expected and business' must conform to survive.

**Current Practise:**

Limited research is available on the cost estimation techniques performed by industry for complex projects, however two surveys were performed in 2000: one on construction projects, the other on IT projects. The construction cost estimation study by Akintoye and Fitzgerald (2000) surveyed 84 UK construction contractors, ranging from small to medium to large, about their estimating practises. They found that the main method for cost estimation was breaking the project into detailed parts and adding up the cost of each item. The next two most popular methods were 'comparison with similar projects based on documented facts', and 'comparison with similar projects based on personal experience'. These can all be classified as experience based models (Akintoye and Fitzgerald 2000). A survey by Moores and Edwards in 1992 of 54 software developing companies found that detailed project planning tools were used by most companies as opposed to cost estimation tools, suggesting that projects were priced based on an analysis of a detailed breakdown of tasks within a project. In the same survey, 91% of software companies cited cost estimation as a problem. Although limited recent literature is available on industry cost estimation practises, historically it can be observed that for complex construction and software projects, cost predictions are chiefly made with a mixture of experience, project comparisons, and an analysis of the project details and tasks. It is also acknowledged that cost estimation is a problem and hypothesised that inexperience and lack of time are chief contributors (Akintoye and Fitzgerald 2000).

Lovallo and Kahneman (2003) tackled the psychology behind the high failure rates of executives in predicting costs of projects such as manufacturing plant construction, mergers and acquisitions, large infrastructure and software development. Their theory stems from Kahneman's work on decision making that won him the nobel prize for economics in 2002. His research argues that humans' 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 downplay or ignore the risk of problems that are out of the team's control. It is for this reason that it does not matter if the project is broken down to the highest level of detail for cost prediction, 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). Lovallo and Kahneman (2003) call the practise of analysing a project based on project details and its unique complexities the 'inside view'. Research has shown that if people are forced to make predictions about their skills after being exposed to an 'outside view', their predictions are significantly more accurate. This outside view can be applied to complex projects, and involves ignoring the details of the current project and instead analysing the outcomes of several projects of a similar description. Mobilising the outside view in this way is called reference class forecasting (Lovallo and Kahneman 2003). One method Lovallo and Kahneman (2003) recommend is to obtain correlation statistics from past similar projects - the correlation between the forecast cost and the actual cost. The correlation for the current project can then be estimated via a statistical model which is used to adjust the forecasted cost made by detailed analysis (the inside view) (Flyvbjerg 2011). The traditional way to think about a project is to focus on the project details. Even though gathering data from similar projects could significantly improve a cost prediction, it is rarely thought of by managers (Flyvbjerg 2011).

The implementation of reference class forecasting began in project management for the first time in 2004, was endorsed by the American Planning Association in 2005, and is now used in some governments and private companies in Europe, South Africa, and Australia. An example of the type of output from a reference class statistical model is a plot showing the relationship between the acceptable chance of cost overrun and the required uplift to the original forecasted cost (Flyvbjerg 2011).



This method was implemented by the British government for a rail project which is still under completion

The visualisation presents a powerful communication tool for influencing decision makers and improving forecast cost accuracy, however access to credible data to a sufficient number of projects can be a challenge (Flyvbjerg 2011). Overall, the theory of optimism bias is a credible explanation for the problem of persistant underestimation of project costs. The accompanying solution of reference class forecasting presents a promising solution given enough accurate data.

The theory of reference class forecasting was first clearly published in 2003, however the idea of predicting project costs from previous data has been the subject of research for many years before that and continues to this day. Although research results have often been promising, industry uptake of the idea is not yet established.

### Overview of Past Reference Class Forecasting Methods and relevant Industry Uptake

In the following section, an overview of the historical attempts to integrate reference class forecasting will be presented followed by published theories for the lack of industry uptake.

#### Existing Models

A substantial amount of research on project cost prediction has been dedicated to the fields of software development and construction projects in particular. These analyses have generally used project information from 15 to 20 large projects. A single dataset may consist of projects from around the world, and each project a different client and project team. The mathematical methods used to predict project costs vary from statistical regressions to deep-learning neural networks (Love, Raymond, and Edwards 2005)(Kim, An, and Kang 2004). Most studies report neural networks are more accurate than linear regression, and generally the studies report positive predictive powers of their developed models.

A McKinsey study outlined a case in the film industry where a motion picture company used reference class forecasting of movie project success (cases were weighted by similarity) to decide which movies should be heavily promoted. This model improved forecasts on financial return by 135% relative to single project analogies (Flyvbjerg, Garbuio, and Lovallo 2014).

Statistical research projects which trial reference forecasting have shown promising results in IT, construction, and film projects. However, there is little evidence that the methods have successfully transitioned into industry.

#### Industry Uptake of Existing Models and Current Practises

Surveys on current industry practise for cost estimation are scarce, however, a survey of the construction industry in the UK in 2000 found that a manager's experience and intuition is still the dominant method for construction cost calculation (Akintoye and Fitzgerald 2000). A study on the IT industry by Jorgensen et al. (2007) cited that the parametric models were not comprehensive enough to be used in industry and also found that expert judgment was not mobilised enough in the process of developing these tools.

Although reference class models have not gained traction in either the software or construction industry (where most of the research to date has focussed), it is worth outlining their current practises in cost estimation and the reasons behind these.

In the construction industry, infrastructure and building projects follow a traditional and well established process, where contractual norms have developed over many decades (Badenfelt 2011). The industry is so large, broad and well established that it has inertia in the way that projects are estimated, and contractually bound. The contractor (builder) first reviews the drawings and if the plans are not detailed enough, they offer a temporary budget price. Once they feel that the drawings are sufficiently detailed, they will offer a fixed price (Badenfelt 2011). However, as the job progresses, if any work is added that was not in the detailed drawings, they maintain the right to charge a variation (Badenfelt 2011). This method is clear and variations are easily justifiable to the client, however cost, labour and time estimates are still regularly under calculated, even from detailed drawings. And for large infrastructure projects, small percentage miscalculations can be exceedingly expensive.

Contractor's practise of pricing from detailed drawings and charging for variations does not translate to other consultants in the construction industry, such as engineers and architects. Consultants must devlop the detailed plans for the contractors, and must therefore offer fixed price contracts based on drawings with little detail. It is much harder for a consultant to charge a variation for a portion of their time spent on something that was not included in the original simple plans. This is despite the real and common instance of budget blow outs due to overspent time (Harris 1999). One method to provide opportunity for a variation is stating a minimum percentage fee of the final building cost, which can overrule the original fixed price contract. If a consultant has spent time beyond their fixed price budget,this may account for some or in the best case, all, of their overspent time.

Construction projects generally are completed more on time and on budget when compared to IT projects as the greatest cost is the project is building materials (Badenfelt 2011). Unlike time, building materials can be accurately calculated from detailed design drawings. Again, this omits data from consultants who are part of the building industry but charge based on their time only.

In contrast to construction, IT projects are a relatively new practise where alternative contractural arangements have been trialled. Through experimentation with contracts, the Agile method was developed. In summary, the method treats both cost and time as fixed quantities for a project, and if any change or additional work can only be accommodated if another, less important requirement is excluded (Badenfelt 2011). Furthermore, the project outcomes are continually adjusted ad revised based on frequent progress review points that assesses unexpected problems. This agile attitude has become ingrained in many IT practises and this flexible upfront arrangement with clients often forges better long term relationships [Cockburn2001].

It is not completely clear why industry uptake of cost estimation tools using the outside view has been weak, and it may be the case that in modern industries such as in the IT industry, problems associated with cost estimation can be innovatively designed out of the contract. Possible reasons that parametric reference class forecasting models developed by researchers have not been incorporated into industry practise include:

* lack of appropriate data
  + if a project is unique, such as a national infrastructure project, it many not feel appropriate to gather sufficient data points from other countries and time periods.
* lack of understanding from industry decision makers - the prediction process was not explained well in marketing of the tools
* models were not developed in collaboration with industry
* does not feel intuitive price from an 'outside view' of a project which disregards small project details
* insufficient time allocated to pricing that makes change burdensome

In fact, a study by Moores and Edwards (1992) concluded that a lack of framework to support the outside view model use and a failure in marketing were the two most common explanations for lack of uptake of cost estimating tools which use statistical techniques in predicting project costs. This is unfortunate as almost all research indicates an outside view analysis of a job substantially improves cost estimation accuracy for fixed price projects.

### Case Study

This research will explore the potential for reference class forecasting to improve cost estimation in the context of an engineering consulting company in the construction industry. In this field, labour is the chief cost and the long traditional history of this industry mandates fixed price projects as the norm. In recent years, the size of losses from unprofitable projects in the case study could equal up to 25-30% of the profits from profitable projects. This demonstrates substantial room for improvement and the possiblity of multitude positive flow on effects.

Cost estimation is currently performed by first carefully reviewing preliminary drawings for a project where time costs will be calculated using a manager's personal experience with similar projects. This will often be cross-checked with a value based off a percentage of the estimated final cost of construction (generally the client budget). No formal mathematical method of comparison to similar projects is performed. This is the case for a number of possible reasons:

* lack of time to interrogate the CRM
* clunky data availability offered by the CRM
* lack of awareness of projects performed by other managers that may have been similar.

Reference class forecasting in this study differs from previous research, in that statistical analysis will be performed on internal past company project data that has been progressively recorded in the company's Customer Relationship Management (CRM) software which was purchased. This presents an opportunity to evaluate reference class forecasting in a unique situation, where the 'reference' projects are internal and over two thousand cases are available.

The CRM data currently stands as an untapped source of information within the case study organisation. It stores a rich variety of data including:

* employee timesheet 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

A statistical model predicting a measure of project financial success will be developed using the available project variables and trialling numerous machine learning algorithms and statistical methods. The company uses well established CRM software which readily provides simple output statistics such as project time-cost vs invoiced amount summaries. However, the analytical capabilities were limited to:

* simple scatter plots and bar charts of the raw data
* summaries such as overall hours spent vs invoiced amount for each project

The company has a wealth of data but limited means to extract insight. This study will exploit the CRM data by performing more sophisticated statistical analysis with the intention of building a predictive model for cost estimation. The benefits of this study in comparison to previous cases are that thousands of past data points are available, when previous studies used on average 15-20 cases and at most 170. This improves the potential for accurate predictions. If successful, the model also has higher potential for managerial uptake as the algorithm will be directly built and trained on company data. Managers can interrogate 'similar' data points and relate to the actual projects or speak to a colleague who was involved. This has the potential to give the 'outside view' stronger influence on the final decision.

### Case Study Limitations

The case study provides a framework to test the value of a consulting company's CRM data in assisting project cost estimates. However, it is important to be aware of limitations when reviewing outcomes and framing the problem. The most obvious limitation is that we are disection a single company in a specific industry, which limits the ability to state conclusions could apply to a broad range of cases. This can be overcome with future studies.

The data set from this case study does not include information that would improve project cost predictions. For example, managers commonly price jobs based on the budget for the entire project, of which the consultant is a minor part. The manager might typically moderate his fee by say 1% of the entire project cost. The value of the entire project has typically not been recorded although there is a space in the project software form for this value. It would be even more beneficial for managers to record the expected total project cost at the beginning of a project versus the conclusion of a completed construction project. Another piece of data that was not recorded was external factors influencing the fixed price quote including competition from similar firms that the client played against the case study company.

In addition to this data that was not typically entered by workers, the dataset also excludes detailed information about the projects which are crucial to the final pricing. This includes details such as preliminary drawings describing the project (square meterage, number of storeys), and project summaries provided by the client to enable a quote. This kind of detailed information will never be recorded in a CRM and in many ways contradicts the purpose of the 'outside view' which intends to present similar projects without being over-influenced by the finer details of a project. Given the lack of detailed information, it is not reasonable for an algorithm to predict the specifc fee of a project. It is much more effective to attempt predict the profitability of a project (ie return per dollar) based on the historical data. This profitability prediction relates to the profitability of the project if the fee was generated using the established method the company used in its historical cases.

Finally, CRM data is entered by employees manually as time passes, jobs are invoiced, and projects are finally completed. Because the data is input by humans, it is susceptible to input error. Many of errors were detectable during data cleaning, however it is possible that undetectable human errors exist in the data.

Although this case study presents an opportunity to test the predictive power held within the organisations' existing CRM data, limitations exist which must be understood and built into the framework for predictions. These include missing data that could have been entered for each project, but historically has not been required, missing data that is very detailed and will never be available, and the potential for a small degree of human error to be embedded within the data.

### Research Motivation Summary

Fixed price cost estimation for complex projects is a difficult task and when performed incorrectly, can have devastating affects on businesses and governments. Many theories have delved into human nature and our tendency to optimistically assess our capabilities with respect to a given task. In this case, the optimism often results in negative business consequences affecting employee morale and the capacity of a company to flourish and grow. Over the past few decades, significant research has been dedicated to creating predictive models that present the 'outside view' of a project, by statistically comparing a new project to numerous similar projects and their characteristics. It has been shown that these models improve the cost estimation, however industry uptake has not been successful. It has proven to be a challenging task to persuade managers to use this so called reference class models in practise and remove themselves from additions based on intricate details of the completed project. This may be due to a number of reasons including lack of understanding, time, and lack of collaboration with industry in development of these models. This research examines the potential for a company's internally generated CRM data to be used in creating a reference class forecasting model. This has the potential to influence the managers' decisions in the company more however its limitations must be understood in order to create a good model.

## Thesis aims

For the reasons explained above, the aim of this research is as follows:

**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 costs before project engagement and other business analyses such as client analysis. Several statistical and machine learning techniques will be 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.

**Hypothesis 2** The predictive model built from Hypothesis 1 can be proven to have a positive impact on the bottom line of the case study business.

## Thesis Contributions - Gaps in literature

* a gap in the study of predicting internal performance in businesses with business clients. The literature has assessed performance of movies, customer analytics but rarely employee performance or business clients. This is particularly relevant to consulting companies that tackle complicated discrete projects with set timelines. Statistical and machine learning models to predict the correct fee for projects according to project type, client, and other characteristics could reduce risk for a range of consulting businesses in setting initial fees.
* A gap in machine learning predictive models which cater for variables with large amounts of missing data. the nature of consulting jobs is quite complex and historically, not all information about a project is possible to obtain. Variables with missing data are alternatively treated as optional bonus variables in a multi-step model as opposed to missing information which must be imputed or deletion of the entire case that contains missing information.

# Literature Review - 25 pages for pHd

This chapter provides an overview of the literature available about current business to business (B2B) cost estimation methods, use cases of CRM data, as well as machine learning algorithms and statistical business prediction models. The most prominent statistical and machine-learning models developed for business applications shall be reviewed along with models applied to the most closely related business cases to this project. Gaps in past research will also be highlighted.

## Project Cost Estimation Methods

This project will tackle the problem of estimating fixed fees for complex consulting projects. It is therefore imperitive to understand what research has been performed to date and how this project contributes to the body of work. The task of estimating fixed fees can be likened to the terms 'cost estimation' or 'effort estimation' in the literature. The bulk of research to date has been performed with project data in either the construction industry or software development. The research into cost estimation in each of these industries is reviewed below, detailing different methods and outcomes. Finally a gap in the literature will be highlighted which this research aims to fill.

### Cost Estimation in the Construction Industry

Research into cost estimation in the construction industry chiefly focuses on predicting the final building cost of construction. Various studies have surveyed popular methods for cost estimation while others have investigated new methods such as building information modelling (BIM) analyisis, case based reasoning, and predictive algorithms. The cost estimation methods of detailed analysis (inside view), case based reasoning, and predictive algorithms shall be reviewed.

**Detailed Analysis**

Detailed analysis refers to the process of an engineer or builder in the construction team carefully reviewing construction drawings to sum the cost of materials, labour, machinery hire, overheads, profit etc. (Akintoye and Fitzgerald 2000). Akintoye and Fitzgerald (2000)'s study based in the UK found that this is the most popular method for pricing amongst contractors. The second most popular method was to perform an additional manual comparison with similar projects and make judgments based on personal experience. Note that this does not involve a mathematical model.

Elfaki, Alatawi, and Abushandi (2014) discusses how much detailed cost estimates for cosntruction projects can vary from estimator (engineer) to estimator, which contributes to the lack of accuracy in predciting final cost of a project. He argues that an engineer's expertise is not documented or measured in any way and therefore their expertise and, in turn, their estimate is prone to subjectivity. Shane et al. (2009) theorises that final cost can be subject to so many different unpredictable parameters, such as weather, unexpected ground conditions, duration or sub-contractor issues that it is almost impossible to achieve an accurate cost prediction manually.

Groundbreaking methods into calculating construction costs by Ma and Liu (2014) mobilise the data from building information modelling, although this method is still technically a detailed analysis of the costs. The idea is to quantify the cost of building a structure directly from a three dimensional model created by the design consultants (engineers and architects). Ma and Liu (2014)'s first trial was to automate the cost calculation of a reinforced concrete structure. His chief development was to program his algorithm to intelligently establish construction techniques for each element as this is not provided by architectural or engineering models. From the construction technique and material, a cost can be determined for each element. This was successful and Ma aim's to further his work so that more construction information can be intelligently obtained from BIM models. A system like Ma's would definitely accelerate the cost estimation process and reduce human error, however it is still an 'inside view' and could be prone to innaccurate estimates in a similar way to a detailed analysis by a person. The detailed view does not 'step back' and take into account set backs in construction that may have affected similar projects in the past. There is potential for this to be built into the same system but this is still in early stages of development.

Surveyed information on the latest techniques for cost estimation is limited, however what is available indicates that cost estimation through detailed analysis (inside view) remains the most prevalent method. This is despite construction having a long history of projects running over time and over budget using the same cost estimation technique (Shane et al. 2009). There is even current research on how to automate this way of cost estimation using BIM. The method of detailed sums shall be compared to the results of studies using parametric algorithms and case based reasoning ('outside view') in the following sections.

**Statistical and Machine Learning Models**

As the problem of cost estimation in construction projects is well recognised, dozens of studies have built prediction models of these projects using machine learning and statistical models. The aim is to test whether a mathematical comparison of many similar projects, using meaningful descriptive variables, can predict the cost of construction. This section will first review the previous studies and most accurate algorithms and finally summarise industry uptake of the methods.

Elfaki, Alatawi, and Abushandi (2014) completed a study on cost estimation techniques and research from 2004 to 2014. It was found that artificial neural networks and support vector machines were the most common machine learning techniques used. He highlighted that these two machine learning techniques deal well with uncertainty, however lack technical justification to the decision maker (known as a black box). Neural networks are also time consuming to train, and must be re-trained and re-tested with each additional piece of data (Kim, An, and Kang 2004). Nevertheless, neural networks and SVM's recieved a significant amount of 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. This is surprising given the rapid uptake of boosted trees over the past decade. In the study, data from 234 school building construction projects in Korea was used. Boosted trees were compared to neural networks, which are one of the most popular algorithms to apply to the problem. Boosted trees were found to predict costs more slightly accurately than neural networks, but not significantly (p value < 0.5). Even though the two algorithms performed similarly, the output from boosted tree models provide insight into the structure of the model such as variable importance and variable partial dependence plots. This additional insight may mean boosted trees are preferable over neural networks because of their potential to engage decision makers.

A notable amount of literature has studied the predictive power of multiple linear regression in the construction cost estimation problem. Often, linear regression is 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). However, even post-1990, many studies focused on linear regression only. This may be because the technique is straightforward, easy to use and widely available in statistical packages (Chan and Park 2005). Out of those that have compared regression to neural networks, the studies have shown that neural networks outperform regression, however other studies show they are approximately equal (Kim, An, and Kang 2004)(Attalla and Hegazy 2003). Dissanayaka and Kumaraswamy (1999) found that neural networks outperformed multiple linear regression models, however regression models could be performed first to determine variable importance and condense the number of variables. It can be argued that boosted trees could perform as well as neural networks while providing additional insights (even more so than regression models) such as variable importance and partial dependence of variables against cost.

Despite neural networks typically outperforming linear regression models, it is intersting to note some of the outcomes of regression studies in cost estimation. Different studies tend to produce a wide range of variables that contribute to different degrees in the models. This is likey due to the nature of construction projects which are multidisciplinary and involve a range of different parties such as the clients, consultants, contractors and suppliers, each of which can have a range of characteristics and degrees of input (Chan and Park 2005). Chan and Park (2005)'s study on 87 building projects in Singapore included special, complex projects. Therefore, variables such as contractor's specialised skills, who the client was (public vs private), the client level of experience, and the contractor's financial management highly influenced final cost of the project. Other studies found variables such as project complexity, duration, team experience, information availablility, site requirements, and labour climate to contribute most (Akintoye and Fitzgerald 2000)(Trost and Oberlender 2003)(Pinto and Slevin, n.d.). As these studies obtained information from questionaires, the available variables also largely depended on the questions in the questionaires, and therefore each study produces a range of different important variables. Despite the differences in qualitative outcomes, the studies all generally report their mean absolute error or a similar metric from the tested models. This is encouraging, however there is a gap in the literature in how to translate the improved errors from the model into an argument for industry to uptake the model as an improvement on their bottom line.

Surveys into construction industry uptake of these algorithmic models are sparse, however a survey by Akintoye and Fitzgerald (2000) showed that parametric estimating methods have not been adopted by contractors, despite the research into this area over 20 years prior. The majority of contractors use experience based methods. Akintoye and Fitzgerald (2000) hypothesises that this could be due to a lack of familiarity and knowledge about the techniques, doubts whether these techniques are applicable to the construction industry, and the availability of sound data to ensure confidence. This presents a clear gap in the literature on how to analyse the models in terms of the goals and limitations that a contractor may experience.

The literature on cost estimation for building projects do not break down the cost estimation task into smaller components, even though a construction project is completed by dozens of consultants and sub-contractors. Many of the consultants, such as Engineers and Architects, produce their own cost estimates for their effort that contributes to the project. Their contracts are often fixed price, with minimal room for movement, which means they wear the risk of over-spending to meet fixed cost contracts (Harris 1999). Furthermore, the existing studies collect data from a wide range of projects that have certain characteristics in common, such as country, or building type and information is obtained through surveys. This leaves a gap in research exploring a single contractor's history of projects, which would be available in their company database.

**case based reasoning**

Case based reasoning (CBR) is a method of estimating by using the results of similar cases. The idea is that in the case of a new project, similar projects are chosen, before generalisations about the data are made. This contrasts to algorithmic models that are built from the entire set of data (Elfaki, Alatawi, and Abushandi 2014). Research into the application of case based reasoning to construction cost estimation began in the 1980's (Kim, An, and Kang 2004). The general procedure involves:

1. Storing a collection of projects with key variable values
2. Once a new project or case arrives, similar cases are retrieved. This can be achieved either algorithmically with a distance function applied to the variables, or manually by users reviewing past cases. The nearest neighbour algorithm has been used to find the distance between cases in previous studies. It is programmed to calculate the euclidean distance in n-dimensional space between cases, where each variable is a dimension (Kumar and Ravi 2007).
3. The cost of the new project is estimated by extrapolating characteristics of the similar cases to the new case. This can also be done either algorithmically or manually by an experienced decision maker.

A study by Kim, An, and Kang (2004) compared CBR to neural networks and linear regression in the construction cost estimation problem. For CBR, the ESTEEM software package was used which algorithmically calculated the similarity of variables, weighted the variables using gradient descent, and deduced the cost using the most similar cases (the specific method for this last step was not explained). The discussion compared mean absolute error rates of the three methods and showed that CBR was more accurate than regression, but less accurate than neural networks in cost estimation. Despite neural networks outperforming CBR, CBR maintained significant advantages over neural networks which have been referenced in other studies as well.

Advantages of CBR include algorithmic efficiency as well as user engagement. In Kim, An, and Kang (2004)'s study, it was noted that CBR models were simple to update with new data in comparison to neural networks, which must be slowly re-trained and re-tuned. New cases also did not need to have every variable complete which can occur often in real, messy data. Neural networks on the other hand require complete data. Another significant advantage of CBR is the ability for users to review chosen similar cases and make sense of the predicted cost, as opposed to neural networks which are a black box (Kim, An, and Kang 2004)(Elfaki, Alatawi, and Abushandi 2014)(Kumar and Ravi 2007). A disadvantage of CBR is the accuracy can be highly dependent on the number of selected cases (Elfaki, Alatawi, and Abushandi 2014). Overall, CBR may provide prediction accuracies between regression and neural networks but its results can be justified by the user - a valuable asset.

CBR can be described as a systematic method of expert judgment, where the decision maker manually compares similar projects from his or her experience (Shepperd, Schofield, and Kitchenham, n.d.). CBR that is performed manually through personal experience, without the assistance of algorithms, was the second most popular method for construction estimation in Akintoye and Fitzgerald (2000)'s survey in the UK. This indicates it is quite an intuitive method for decision maker's to cross check their detailed cost estimate, although an algorithmic version of CBR has not been taken up by industry.

**Summary**

In summary, research into cost estimation in the construction industry has been undertaken using a variety of methods ranging from detailed manual analysis, algorithmic models, and case based reasoning. Neural networks generally outperformed all other systematic methods, such as regression and case based reasoning, but had a significant disadvantage of being a black box. Alternative methods such as boosted trees have the potential to perform as well as neural networks and provide insight into the structure of the algorithm. This is true for CBR as well, although it has not been shown to predict as accurately as neural networks. There exists a gap in the research in ways to intelligently combine CBR with machine learning methods that could predict well, provide insight to the structure, and engage the user in reviewing similar projects. Also, as stated previously, there exists a gap in the application of cost estimating in the construction industry to smaller components of the project, such as consulting and contracting companies who each face a similar problem in estimating their fixed fee. These companies could mobilise information in their internal project databases, as opposed to available literature which surveys dozens of companies involved in selected construction projects.

### Effort Estimation in the IT Industry

Like the construction industry, the software industry faces the challenging task of estimating the cost of a project before full engagement with their client. However, unlike construction projects, the major cost associated with the project is *effort* as opposed to the cost of building materials. A parallel can be drawn between software project effort estimation research and consultants effort in the construction industry, which was a gap in the research of the previous section.

Effort estimation is indeed a problem, and was demonstrated by a review of 6 surveys of IT projects between 1984 and 1994 by Moløkken and Jørgensen (n.d.). It revealed that 60-80% of IT projects encounter effort and/or schedule overruns, where the average overrun was by 30 to 40%. A survey of software managers asked whether they saw effort estimation as a problem, and 91% responded 'yes', while only 9% answered 'no' (Moores and Edwards 1992). The most popular response for the supposed reason for cost overruns was over-optimistic estimates (51%), which falls squarely within the theory of optimism bias, or the 'inside view' (Flyvbjerg 2011). Given this widespread problem in the industry, plenty of research has been dedicated to the problem of effort prediction, with methods similar to the construction industry literature: detailsed analysis, algorithmic modelling, and case based reasoning.

**Detailed Analysis**

Similar to the construction industry, expert judgment or detailed analysis, is the most widely practised method for effort estimation (Shepperd, Schofield, and Kitchenham, n.d.)(Moløkken and Jørgensen, n.d.). This is despite many years of research being dedicated to developing algorithmic models and industry tools that, in the research context, outperform expert judgment. A survey by Moores and Edwards (1992) found that software tools assisting a software project were in significant use, but only the project planning tools, not the cost estimating tools. This suggests a mental link for project managers to price jobs based on *details* of the software project as opposed to taking an outside view. Another study by Heemstra (1992) found that there was no evidence estimation accuracy improved when estimation tools were used. Bergeron and St-Arnaud (1992) similarly found jobs that used algorithmic models were actually associated with less accurate estimtes. However, this trend may be coincidental due to a lack of cases where estimation tools were used (Moløkken and Jørgensen, n.d.). The scarcity of real project evidence that algorithmic estimation tools improve estimations and an unintuitive mental jump could be reasons contributing to detailed expert analysis remaining the most widespread technique for effort estimation.

**Algorithmic methods**

Algorithmic methods to predict effort for IT project costs has been a popular subject of research. Regression models have been the most substantially tested followed by neural networks (Finnie, Wittig, and Desharnais 1997). This section will assess the details and performance of the algorithmic models, important variables, applications and finally gaps in the literature.

Multiple studies have shown that neural networks definitively outperform regression models in effort estimation, although regression is the most popular method in the literature (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 models accuracy: size (Shepperd, Schofield, and Kitchenham, n.d.)(Finnie, Wittig, and Desharnais 1997)(Pai, McFall, and Subramanian 2013). Size can refer to the expected number of lines of code or function point. Function point refers to the amount of business functionality expected from an information system product. This is easier to guess correctly at the beginning of a project, and therefore leads to more accurate predictions (Finnie, Wittig, and Desharnais 1997). One study used a single variable approach (size) with a linear coefficient and an exponential coefficient as follows:

This model was compared to case based reasoning, and case based reasoning outperformed it (Shepperd, Schofield, and Kitchenham, n.d.).

There are disadvantages to algorithmic methods, similar to the construction industry. First, there are often not enough cases to create a good model in the software industry, particularly if the cases must be from within the company (Finnie, Wittig, and Desharnais 1997)(Pai, McFall, and Subramanian 2013). A study by Mendes and Kitchenham (n.d.) demonstrated using 67 web projects that cross-company models were significantly less accurate than a within company model, so it is in the company's interest to create an in-house model. This means that effort estimation tools should use in-house data and must be adapted by statistical experts to each company (Shepperd, Schofield, and Kitchenham, n.d.).

Algorithmic estimation tools have not been successfully adopted by industry. This may be because no model has proved to be outstandingly successful at consistently predicting required effort (Finnie, Wittig, and Desharnais 1997). Also, neural networks (the most accurate algorithm) lack the ability to capably explain its results and instill trust in decision makers (Finnie, Wittig, and Desharnais 1997).

The literature heavily focuses on effort estimation for the software industry. This leaves a gap in studies for effort estimation for consulting companies in other industries such as the construction industry. Construction projects do not have a simple 'project size' or function point variable and would need to utilise other variables available.

**Case Based Reasoning**

In the IT industry, applications of case based reasoning to effort estimation have been researched in a similar manner to the construction industry. The procedure is described as a algorithmicallyfinding similar projects to a new case using measurements of euclidean distance in n-dimensional space, where each dimension corresponds to a variable (Shepperd, Schofield, and Kitchenham, n.d.). The similar cases are then used to predict effort for the new case, and in Shepperd, Schofield, and Kitchenham (n.d.)'s study, linear regression was used for this final step.

In contrast to the construction industry, the literature in software effort estimation found CBR to perform approximately equally to algorithmic methods such as neural networks (Shepperd, Schofield, and Kitchenham, n.d.)(Finnie, Wittig, and Desharnais 1997). It was also highlighted that CBR is intuitively similar to how an expert thinks about pricing projects and provides insight into the prediction that neural networks do not (Finnie, Wittig, and Desharnais 1997). Research into CBR in the IT industry found similar results to the construction industry except that CBR performed as well as algorithmic methods.

**Summary**

Research into effort estimation in the software industry has followed a similar path to the cost prediction of construction projects. Expert judgment via detailed analysis is still the predominant method for effort estimation. Case based reasoning and algorithmic methods such as neural networks have found some success in effort estimation (at similar levels) but have not been successfully adopted by industry. It was noted that case based reasoning resonated more with decision makers because its effort prediction could be justified by the model. This visibility of the model prediction structure is very important for future work if the results are to be translated to industry. Effort estimation remains a problem, however agile methods of project management and effort management are changing the contractual approach of consulting in the software industry (Badenfelt 2011). This innovative method keeps delivery times short (and therefore predictable effort) and defines projects as something with a fixed cost and time. This means if a change occurs, other less important elements must be discarded maintain the fixed amount. Software development is a relatively new field in comparison to the construction industry and is able to adopt this new model relatively quickly across the industry. However, effort estimation for consulting companies in the construction industry are faced with a long history of tradition that is much harder to tackle. This leaves a gap in research in algorithimic models that predict effort for consulting companies in the construction industry. From previous literature, the models should allow decision makers to interrogate the structure of the prediction and be based on internal company data.

## Statistical and machine learning methods

### Introduction

The aim of this project is to use statistical techniques to model the profitability of projects. In order to optimise the model, several statistical and machine learning techniques will be tested, and compared and it is therefore prudent to review the range of statistical and machine learning techniques that have been successfully applied in the literature. These vary from simple methods such as Linear Regression to complex, deep learning such as Neural Networks. The previous section highlighted the use of linear regression, neural networks, SVM and in one case boosted trees for prediction, however other business problems have utilised a wider range of methods. These include Naiive Bayes, Random Forests, and machine learned Bayesian Networks. The following section will present each method's advantages and disadvatanges and an example of a past successful application.

### Linear Regression

Linear regression is one of the simplest and most popular statistical prediction methods (Hastie, Tibshirani, and Friedman 2009). An equation for a line is learned, which is defined by a constant and a coefficient times the value of each variable. The coefficients are chosen to minimise the residual sum of squares (least squared) (Hastie, Tibshirani, and Friedman 2009). The reason this method is popular is that it is simple, provide stable predictions, and can be adapted for categorical variables (Seng and Chen 2010). The integration of categorical variables is called Anova, and the algorithm compares group mean variances within the categorical variable.

There are several drawbacks to linear regression. For example assumptions must be made about the structure of the data. Predictor varialbes are assumed to be linearly related to the response variable and the varialbes are normally distributed. Many real world phenomena do not correspond to these assumptions which can make it difficult, sometimes impossible, to produce useful results (Hastie, Tibshirani, and Friedman 2009)(Putler and Krider 2012). Furthermore, response variables must be continous numeric variables and if two or more predictor variables are correlated, the values of their coefficients are highly unstable. On the other hand, correlated variables do not impact the predictive accuracy of the model. A final drawback is that the variable coefficients cannot be used to compare variable importance unless the varialbes have all been standardised to one another (Putler and Krider 2012). This limits the amount of insight that can be drawn from a linear regression model.

Despite the limitations of linear regression models, they have widespread business applications due to their simplicity and ease of application. Regression is used to predict risk in the finance and insurance industries, predict who to target in marketing exercises, and predict consumption spending in the field of economics (Harrell 2013). Due to its simplicity and widespread application, linear regression is a good starting point for an analysis and benchmark for performance against complex models.

### Logistic Regression

The method behind linear regression can be adapted to predict a binary response variable (1 or 0). A predictive linear equation is computed so that the predictor variables are combined with coefficients to give an output between 0 and 1. This is then transformed into the log odds to mediate extreme variable input values. The result is a probability an input case will result in a 0 or 1 (i.e. success or failure) (Moore and McCabe 1989). Similar pros and cons exist for logistic regression as linear regression, and it too can be used as a benchmark to compare other binary predictive models due to its speed and simplicity. Many business problems have binary response variables, such as yes/no, male/female, buy/do not buy, success/failure, or survive/death (Moore and McCabe 1989).

### Naive Bayes

Naive Bayes prediction method makes conditional independence assumptions about the predictor variables in order to greatly simplify probability calculations for the response variable (the response variable must be categorical). Specifically, it is assumed that each variable is conditionally independent given the respose class alone, without considering other predictor variables. Mathematically this means that each variable's probability contribution can be simply calculated by looking at the probability of a response class given a predictor value by itself. The predictor variables probability estimates are combined for each possible response class, and the response class with the highest probability is chosen (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 very well for real world tasks because the assumption of independence does not damage predictions significantly. This is because if multiple varialbes are related, the variables independently still direct the prediction in the correct direction. The correlated variables will double or triple their emphasis on the predicted response variable, however this does not necessarily impact classification accuracy as the final step is to simply choose the class with the *highest* probability (Provost and Fawcett 2013). This is fine for ranking. On the other hand, the output probabilities are not realisitic, and the values themselves should not be used (Caruana and Niculescu-Mizil, n.d.). Another disadvantage is that in order to calculate the class probability from a numeric variable, a distribution must be assumed (often gaussian) which the numeric data may not fit into neatly (Caruana and Niculescu-Mizil, n.d.). Finally, Naive Bayes classifers are known to not perform well in binary classification (Caruana and Niculescu-Mizil, n.d.).

A real world success story for Naive Bayes classifiers hinges on their ability to be refined incrementally with each piece of new data, as opposed to re-calibrating the entire model. Naive Bayes has been used in complex spam detection systems where new spam emails or toxic text themes can be quickly added to the filtering model. The Naive Bayes method provides a good benchmark to compare against more complex models that should outperform it (Caruana and Niculescu-Mizil, n.d.).

### Decision Trees/ Ensemble Trees

Decision trees are one of the simplest and most intuitive machine learning methods. There are several brands of basic decision tree algorithms including ID3, C4.5, CART, and CHAID with CART and C4.5 being the most popular (**???**). Ravi Kumar and Ravi (Kumar and Ravi 2007) recommend the CART algorithm as it is capable of solving both classification and regression problems whereas the remaining decision trees solve classification problems only. Single decision trees have the risk of over-fitting the data and low predictive accuracy but have favourable characteristics such as providing intuitive rules that a decision maker can follow in real-life scenarios. In addition their results do not depend heavily on the skill of the data analyst (**???**)(Putler and Krider 2012).

To tackle the problem of low predictive accuracy and instability, ensemble tree methods were pioneered in the 1990's with success (Breiman 1996). Three examples of these ensemble decision tree methods are bagging, boosting, and random forests. Bagging creates multiple trees by sampling a different training data set for each tree and then combining tree results (Breiman 1996). Random forests is an advanced form of bagging in that multiple trees are sampled from the same training data. However, when creating the tree for each sample dataset, a random subset of attributes is used to determine each split. This prevents the trees from becoming too correlated (Breiman 2001a). The boosted decision tree approach combines gradient descent with a series of decision trees. Each tree is limited to a certain depth to maintain simplicity, and each tree models the residuals (or errors) from the preceding tree. 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).

Ensemble decision trees have a different foundation to more traditional statistical prediction methods such as regression. For this reason, there are some advantages and disadvantages. The progression of binary splits means non-linear feature interactions can be captured and the predictor variables do not need to be transformed as no assumptions are made about the data's statistical distributions (Louppe and Prettenhofer 2014)(Radenkovic 2010). Other advantages are that the ensemble methods are very fast in comparison to SVM's and neural networks but perform just as well and provide insights such as variable importance and variable relationships (Sealfon and Gymrek 2012). Random forests in particular has the benefit of not being affected by noise in the data as well as the capacity to handle a large number of variables (Sealfon and Gymrek 2012). Caruana (Caruana and Niculescu-Mizil, n.d.) 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. Clearly, ensemble tree methods can compete with high level machine learning algorithms.

Along with the many advantages of trees and ensemble trees, it is important to be aware of their limitations. As trees are not built on a probabilistic framework, results can not be provided in this framework. For example, confidence intervals for predictions are not available for standard ensemble methods (Louppe and Prettenhofer 2014). Also, variable importance tables provided by random forests and boosted trees can be biased towards variables with many categories and another type of ensemble method, conditional forests, should be used (Radenkovic 2010). The methods can also be prone to overfitting if not handled carefully. Ensemble trees have many advantages particularly predictive power, but it is important to recognise that the best method for a problem often depends on the type of problem (Louppe and Prettenhofer 2014)(Caruana and Niculescu-Mizil, n.d.).

### 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 DAG (directed acyclic graph) and is capable of linking the conditional dependency between any variable to another variable or via other variables which are represetned by arcs (Heckerman 1998). This kind of conditional relationship is bayesian. Bayesian networks have found success in combining deterministic models with observation data as well as expert knowledge as the statistical relationships between variables can be determined from different places and manually entered into a single model (Kragt 2009).

Bayesian networks are excellent at representing knowledge from different sources and types, and visuallising them in a way that communicates results well to decision makers (Kragt 2009). Therefore bayesian networks become an excellent decision support tool. Network relatitonships can be machine learned, however this is not widely included in the suite of machine learning methods and is referred to more as a way to model complex networks. Drawbacks of bayesian networks include their inability to handle continuous numeric variables (which must be discretised). This means that the varialbe relationships are categorical (Kragt 2009). Popular applications of bayesian networks are modelling uncertainty in natural resource management and modelling complex business network structures such as airports (Wu and Mengersen 2013).

### SVM's and Neural Networks

Neural networks were inspired by biological neural networks of the human nervous system and require a lot of training data. They are very flexible in the relationships they can mimic and are not greatly dependent on the skills of the analyst (Putler & Krider, 2012). A disadvantage is that although the results are accurate, the algorithm is a black box - meaning the data is input and the results output without providing the user insight into the methodology. Neural networks can perform high level tasks as a result of deep learning such as hand writing recognition, vehicle control, face recognition, and cancer detection (Haykin and Network 2004).

The simplest type of Support vector machines (SVM's) calculate the linear boundary between two categories in feature space so that the category of future cases may be determined depending on which side of the boundary they fall. The location of the boundary is determined by the position of the widest clear margin that can be drawn between the two categories, allowing a certain number of misclassifications. This concept can be adapted to non linear boundaries by performing kernel transformations on prior variables as well as regression problems by introducing loss functions (Auria and Moro 2008). Advantages of SVM's include the ability to model non linear relationships in data, and that no assumptions about the distributions of the data must be made as the model is non-probabilistic. They are also known to be robust over different samples and perform well in high-dimensional feature space (Auria and Moro 2008). On the other hand, the output is in the form of distance to the boundary (as opposed to probability) and results are not transparent (similar to neural networks) (Caruana and Niculescu-Mizil, n.d.). Also, SVM's are known to be very slow to train and therefore often not suitable for industry purposes (Auria and Moro 2008). Because SVM's perform well in high-dimensional space and can form non-linear rules they have been applied to protein classification (medicine) and text and image recognition (Byun and Lee 2002).

### Summary

This section reviewed the advantages and disadvantages of statistical and machine learning techniques that have been successfully applied to business problems. These include linear regression, logistic regression, naive bayes, decision trees and ensemble trees, bayesian networks, neural networks, and SVM's. Linear/logistic regression and naive bayes are computationally very fast with a simple concept that make broad assumptions about the data, but work well as baseline models against which to compare complex models. In certain cases these simpler models are significantly outperformed by complex models however it is important to compare both. Neural networks and SVM's are excellent machine learning predictors which can perform well with messy data and no assumptions, however they are slow and do not provide insight into the results of analysis. This is not appropriate for the profitability prediction problem in this project because communication of the reasoning behind predictions is crucial for uptake of the model by decision makers.

On the other hand, ensemble tree methods can perform as well as neural networks and SVM's and make no assumptions about the structure of the data, but provide insights into the model such as variable importance and variable relationships. Bayesian networks are excellent for presenting the structure and reasoning behind the predictive model, however are very slow to train and have generally not been included in the machine learning literature as much as the other methods. This may be due to low awareness of learned bayesian networks or lower performance. In conclusion, simple models such as linear/logistic regression and naive bayes are good baseline models to implement, while ensemble tree methods and bayesian networks may handle messy data better than the simpler models while still providing insight to decision makers.

## Statistical and Machine Learning Methods Applied to Business Problems

The aim of this research project is to create a predictive model for the profitability of a company's consulting projects using their internal CRM data. The literature pertaining closely to the cost prediction topic has been reviewed in section 2.1, however it is relevant to broaden the assessessment of the literature on advanced machine learning techniques that have been applied to general business problems. The simpler statistical methods have been covered already, so the focus of this sections is on the topic of machine learning prediction methods for business decision-makers. Popular applied topics include predicting stock fluctuation, customer churn analysis, fraud prediction, customer classification, market segment analysis, product success prediction, and recommendation systems (Seng & Chen, 2010). This section will first discuss the use and merits of the employed more advanced machine learning methods, followed by a case study of employee-churn prediction.

### Summary of Advanced Methods

According to Breiman (Breiman 2001b), simple theoretical models such as linear regression are informative but often provide unsatisfactory accuracy for modelling real life data which is messy and does not necessarily align with the rules of statistical theory. As this project will address a problem with a large quantity of messy data, it is worthwhile reviewing the algorithmic machine-learning predictive models that have developed rapidly in recent years. They have found success because they are able to sometimes fit real-life data with more accuracy than theoretical models (Breiman 2001b). Ravi Kumar and Ravi (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. The most widely used model was neural networks, however logistic and linear regression, decision trees, SVM’s, discriminant analysis (DA), and statistical clustering techniques (such as K Nearest Neighbour) were also popular. Supporting Breiman’s (Breiman 2001b) statements, it was found that DA and linear regression techniques were not preferred due to their low accuracy. The overall assessment was that SVM’s outperformed Neural networks (back propagation neural networks were used most extensively), which sometimes outperformed decision trees, and the rest of the methods were generally inferior. Although SVM’s performed the best, as discussed, they are often extremely complex and slow, requiring a high memory (Kumar and Ravi 2007). In support of SVM predictive performance, a study by Davenport and Harris (Davenport and Harris 2007) concluded that statisticians experienced in predictive machine learning algorithms generally acknowledge that SVM’s yield the highest prediction accuracy compared to other machine learning algorithms.

As SVM’s are often computationally too expensive and advanced for a smaller scale business applications, such as this research project, it is useful to compare neural networks and decision trees which are simpler than SVM’s but are also capable of achieving very accurate predictive models. Ravi Kumar and Ravi (Kumar and Ravi 2007) found that neural networks and decision trees were both capable of out-performing the other, depending on the context – which is limited to bankruptcy prediction in Kumar’s study. Both have advantages and disadvantages to consider (refer section 2.2), however the most notable drawback for neural networks is the difficulty in interpreting the algorithms' explanation of human behaviour – a valuable feature in business problems (Tsai and Chiou 2009). On the other hand, trees are known to be intuitive to understand and have been argued to function similarly to how the human mind thinks (Carrizosa, Martin-Barragan, and Morales 2010).

Ravi Kumar and Ravi's (Kumar and Ravi 2007) study concludes that ensemble techniques, which refers to combinations of two completely different algorithms, can often outperform individual techniques. For example, combining the contrasting advantages of neural networks and decision trees is a worthwhile ensemble technique. A paper studying financial earnings management prediction combined these two methods by running them one after another. The neural network was run first to create a high rate of prediction accuracy via its complex non-linear learning properties (Tsai and Chiou 2009). Then, to resolve the lack of explanatory qualities that neural networks are notorious for, decision trees were employed. The 81% of cases that were correctly predicted by the neural network were used to generate decision trees, and in turn useful decision rules (Tsai and Chiou 2009). This strategy is a promising way to benefit from the strengths of well performing complementary techniques.

### Employee Churn Case Study

The literature on predictive business models is generally limited to studies on product sales, customer behaviour, or overall performance of large businesses. There are far fewer studies that assist with internal business decisions dealing with employees, teams, and delivering projects as is the case with this project. An exception is Saradhi and Vijaya’s (Saradhi and Palshikar 2011) study on employee churn, where ‘churn’ refers to the number of individuals moving out of a group within a certain time. Saradhi and Vijaya applied popular customer churn predictive models to employee churn - a novel application that focussed internally on employees rather than externally on customers. The project chose three machine learning classification algorithms that are normally used in customer churn and it was proposed that the variables and model behaviour of employee churn could be related to customer churn. The associated costs of losing customers and finding new customers can also be correlated to the costs of losing staff and hiring new staff.

Naive Bayes, Random forests (an ensemble decision tree method) and SVM’s were built. A comparison study showed that Naïve Bayes is outperformed by random forests which match Saradhi and Vijaya’s prediction (Caruana and Niculescu-Mizil, n.d.). The SVM model actually far out performed random forests and Naive Bayes by achieving 81% vs 51% and 55% correct predictions respectively. This was attributed to the ability to introduce class penalties in SVMs whereas the other two methods are limited by class imbalance (Saradhi and Palshikar 2011). This means that SVM’s had a superior method of balancing the importance of the variables in the data set which was skewed by a heavy majority of employees who kept their job versus the minority who left (25% churn). It should be noted that the boosted tree ensemble method performs a similar task of weighting misclassified cases or outliers and it would have been worthwhile comparing this method to SVMs. Particularly because boosted trees provide insight into the predictive model which could aid uptake of the model with decisoin makers. However, it is noteworthy that their study proved that predictive techniques used on customres could be translated to predicting internal behaviour of employees.

### Gap

The limited literature on analysis similar to Sardhi and Vijaya’s Employee Churn study (Saradhi and Palshikar 2011) highlights a gap in the application of predictive techniques to model internal performance in businesses. This is particularly relevant to consulting companies that tackle complicated discrete projects with set timelines. In this project, internal consulting project data will be used to predict a new project performance in terms of internal employee structure as well as client characteristics. A review of the literature in machine learning applied to business problems revealed that SVM's have found to be the most accurate technique followed by neural networks and ensemble tree methods. As discussed in section 2.2, it was also noted that neural networks and SVM's do not provide insight into the model's predictions to the decision maker. A gap exists in applying these cutting edge machine learning techniques to internal business decision-making.

## Conclusion

The literature pertaining to the aim of this research project, cost prediction using internal consulting CRM data, has been extensively reviewed. The problem of cost prediction has been addressed in many studies beforehand which have generally been limited to prediction of construction costs for buildings and infrastructure and effort estimation for IT projects. The prediction models consisted of case based reasoning, linear regression, neural networks and in one or two instances an ensemble tree or SVM. This leaves a gap in the literature in effort estimation for consulting companies in the construction industry as well as trialling a broader range of prediction methods.

An exhaustive list of machine learning and statistical methods were reviewed theoretically as well as practically in the broader scope of business problem prediction. It was determined that bayesian networks and ensemble tree methods have potential to perform estimation as well as complex algorithms such as neural networks while providing the additional benefit of insights into the reasoning behind model predictions.

# Chapter 3 Method - 2 to 8 pages

The aim of this project is to determine whether the profitability of projects for consulting businesses can be predicted using statistical and machine learning techniques. As stated in the literature review, a gap in research exists for testing whether internal CRM data from a company in the construction industry could predict project profitability. This section will explain the chosen methodology for testing this aim using a case study Engineering consulting company that charges for their time spent on project delivery. Once the predicition potential has been assessed, a secondary aim is to determine the positive impact the predictive model could have on the overall profitability of the company.

## Obtaining data

An assessment of the literature revealed that the predictive power of internal timesheet data has not yet been tested yet for project profitability for consulting companies in the construction industry. In addition, this is the sole source of project data available for these types of businesses besides interviewing project managers and reviewing the account records of invoicing versus employee costs. As the internal CRM for the case study company stores a detailed account of time spent on each project along and client information for twelve years, it was clear that this source of data could reveal the most and was most suitable to statistical analysis. The complete twelve years of project timesheets and invoices was extracted via the CRM software interface, describing 4169 projects. The projects varied from total invoiced amounts of $500 to over $1,000,000 across four internal disciplines in the company. At the time of data extraction, it was not clear whether the full twelve years were necessary and relevant to predicting current project profitability but this was to be determined later in the analysis.

Once the data was extracted from the CRM, an employee from the case study company performed the task of de-identification. This is an important step before research commences as the privacy of project employees, clients, and project names must be removed for ethical reasons. In the case of employee names, professional titles replaced the names, and for client codes replaced client names.

## Cleaning and variable engineering

The obtained data contained a rich source of project information and a lengthy process of cleaning the data and engineering useful variables was required. The data was initially in three types of datasets extracted from the same CRM software:

* Invoicing data
* Timesheet data
* Project summary data

These three sources needed to be compiled into a single dataset that detailed one project per row as this is the structure required for predictive analytic methods when predicting the overall profitability of a project. Before compilation could begin however, thorough cleaning is required where the data was plotted and statistically assessed so that outliers could be visually or statistically discovered. Outliers were then investigated for data entry errors. Many such errors were encountered. One type of error for example was a user entering a 123 kilometer drive in a car as 123 hours. Once these errors were detected, they were discussed with the company directors and appropriate corrective action was taken.

### Variable Engineering

Although dozens of variables were available from the initial dataset, it was prudent to engineer further descriptive variables, particularly in the invoicing dataset and timesheet dataset where the information needed to be compiled into a single row per project. All variables would eventually be tested for variable importance and predictive power with respect to project profitability. Examples of these engineered variables include:

Timesheet Data

* percent of hours performed by each professional role for the entire project
* timespan of the entered project hours
* percent of hours performed by 'profesional' employees as opposed to 'technical'
* total cost of employee hours per project
* total cost of external subcontractors or disbursements per project
* total number of users that entered hours on each project
* mean hours per day entered on a project
* number of disciplines active in a project

Invoicing data

* total amount invoiced and renumerated per project
* mean invoice size per project
* mean invoice per client
* invoice frequency per project
* client invoice frequency

Project Data

* text analysis of project descriptions detected a list of key words that could classify projects into 16 categories. This key word analysis was done in conjunction with a company employee and the resulting classifications were reviewed.
* number of projects completed with each client and client contact

Once the three data sources were combined, further variables were engineered:

* project profit
* return per dollar for each project (how much was spent on expenses for every dollar earned?)

Besides the engineered variables, additional variables included in the project summary dataset were client industry, discipline, job description, and post code. Once numerous potentially important variables were engineered, the most important ones could be narrowed down. This improves the accuracy of a model because unnecessary or irrelevant variables add noise to the prediction of target values. Variable selection also increases computational efficiency by reducing the number of calculations and improves understanding of the prediction structure (Weisberg 2005).

## Variable Selection

### Outlier Deletion

Before determining which variables contribute most to the target value, return per dollar, outliers were deleted. Specifically, cases with extreme return per dollar values were investigated as special scenarios may have contributed to these values and the model is not intended to predict special scenarios. Outliers were initially defined as return per dollar values sitting outside 1.5x the interquartile range (IQR) from the upper and lower quartiles as intially practised by Tukey, the inventor of the box plot (Tukey 1977). The outlier projects determined by this method were then reviewed with an employee from the case study company. Based on their domain knowledge and assessment of the outlier cases, cut-off values for the range of return per dollar values were determined. Outliers were projects with return per dollar values greater than 3 or below -2.

### Variable Selection Methods

Three methods of variable selection were trialled and compared to test whether the different methods highlighted different variables as important. These were Analysis of Variance (ANOVA), conditional inferece forests (cforests), and random forests. It was expected that random forests would highlight variables that had more categories as more important which is a well known bias; it is also a bias that cforests have overcome (Strobl et al. 2007). ANOVA is based on linear theory which contrasts conditional forest and random forests more free-form structure. Therefore, it is important to compare the contrasting methods' assessment of variable importance.

When running variable importance tests, the dependent variable must be clearly defined. For this problem, the dependent variable can be numeric or categorical: whether a project was profitable (binary classification) or *how* profitable a project was (regression). The regression model was chosen as a variable important for predicting the degree of profitability would also predict profitable vs not profitable.

#### ANOVA

ANOVA is closely related to linear fit models but incorporates the analysis of differences in group means (i.e. categorical variables) (Lunney 1970). The method is simple and fast but assumes variables are normally distributed and linearlly related. The variables must therefore be normalised before analysis, however it cannot be assumed that each variable can be perfectly normalised. A disadvantage of this method is that linear relationships and normal variables do not necessarily represent real world data in the best way (Breiman 2001b).

In order to compare variable importance, the output of the model is assessed which includes the linear coefficients for each variable and the p-values for the coefficients (Markham, n.d.). The magnitude of the coefficient does not indicate importance relative to the other variables because the magintude is directly dependent on the values within the variable. This leaves the p-values which are the result of a test for whether there is a significant relationship between the variable and the target variable. This is a good indication of whether the variable has a relationship with the target variable, however doesnt necessarily rank importance. If a variable is measured more precisely, it will have a smaller p value, whereas a variable measured roughly will have a higher p value. This does not necessarily make the more precisely measure variable more important than the other. The p test does however distinguish between variables that have a statistical relationship to the target variable and ones that do not. With regards to p-values it is also important to keep in mind that, with a dozen variables, chance alone can produce a variable with a pvalue < 0.001 7% of the time (Rice 1989).

The p-value output from ANOVA models give a good indication of which variables have a statistically significant relationship with the target variable. However, the values should not be used to rank the variables against one another, and significant p-values can occur by chance.

#### Random Forest

The random forest algorithm can produce a permutation variable importance for each variable as part of its output. Variable importance is represented as a score, and to do this a permutation and calcuations are performed on each variable. More specifically, consider a single covariate. The values of this variable in the out of bag sample are randomly permuted (reordered) for each tree - a technique that mimics the absence of that variable. The trees in the random forest are then run again. Variable importance can now be derived by comparing the results of the forest where the covariate has been permuted to the original random forest. If the accuracy of the premuted forest is much less than the original forest, then that variable is important. Importance scores are calculated as the mean decrease in accuracy over all trees for that permuted variable in the random forest (Breiman and Cutler 2005).

The random forest variable importance provides valuable insight because it addresses the impact of each predictor variable individually as well as in multivariate interactions with other predictor variables. It handles uncorrelated variables well if the samples for the trees are subsampled without replacement (Strobl et al. 2007). However, a severe disadvantage of random forest variable importance is that it is not reliable for variable selection where potential predictor variables vary in their scale of measurement or their number of categories. The importance of correlated predictors is overestimated and the algorithm tends to select variables that have many possible splits or many missing values (Strobl et al. 2007). Overall, the random forest permutation importance is useful as it is able to compare variables in a machine learning environment and can include multivariate interactions, but the current random forest importance measure overestimates the importance of variables with many categories and some numeric variables.

#### Cforests

Cforests are an alterantive ensemble tree method to random forests that overcome the aforementioned shortfalls of random forests such as overestimating the importance of variables with many categories and numeric variables. It is built from ctrees that are based on a conditional inference framework. The key difference between ctrees and standard decision trees are that a significance test procedure for splitting instead of an purity meausre such as the Gini coefficient, which is centered around information gain.

The details behind the significance test procedure of a ctree are as follows. Each variable is permuted in every possible way, and a correlation value is calucated between the tested variable and target variable, for each permutation. The unchanged variable correlation is then compared with the correlation values for all permutations of that variable. From this, a p-value for the true correlation value can be calculated. The predictor variable with the lowest p-value is selected as the splitting variable (Hothorn, Hornik, and Zeileis 2006). Using an ensemble forest of these trees, permutation is then calculated in the same fashion as random forests, via permuting a variable, re-running the forest and comparing the decrease in accuracy. Several sources recommend cforests over other machine learning variable importance methods due to its unbiased qualities (Strobl et al. 2007)(Strobl, Hothorn, and Zeileis 2009)(Hothorn, Hornik, and Zeileis 2006).

#### Variable Selection Summary

Limiting the predictive model to a concise set of meaningful variables reduces noise and improves predictions. Less variables means that a simpler model is being used for prediction which is easier to understand for stakeholders (Weisberg 2005). For these reasons, a subset of the most important variables were chosen before modelling began. This entailed first eliminating outliers, then comparing important variables from ANOVA as well as random forests and cforests. The literature implies that cforests would suit the data for this project as it is unbiased and not limited to linear theory (Strobl et al. 2007).

## Model selection

Previous studies predicting project profitability in software and construction projects were largely limited to case-based reasoning, regression and neural networks. There is a need to test more sophisticated machine learning algorithms that are as powerful as neural networks but provide insight into the reasoning behind predictions. Ensemble trees (boosted trees and random forests) and bayesian networks fit these criteria. Regression and naiive bayes models were also included as simple baseline models. Complex models should be measured against simple models that can be computed at a fraction of the cost. The complete list of models for this study are:

* Bayesian network
* Random forest
* Gradient boosted trees
* Regression - baseline model
* Naiive bayes - baseline model

Some models have limitations and required additional data processing steps such as normalising numeric variables (Bayesian Networks and Regression) and discretising continuous variables (Bayesian Networks). Discretising numeric variables was performed by generating a hierarchical dendrogram of each variable to visualise the clusters. Between four and six clusters were chosen and summarised to find the maximum and minimum values within each cluster. Then, each job was assigned a cluster label to replace a numeric variable. It was decided that for timespans, the discretised variable should be applied to all models because it is easier for a manager to predict a time category than the number of days. For example, a small job could be confidently assigned to less than three weeks and a large job could be assigned to 1.5 - 3 years.

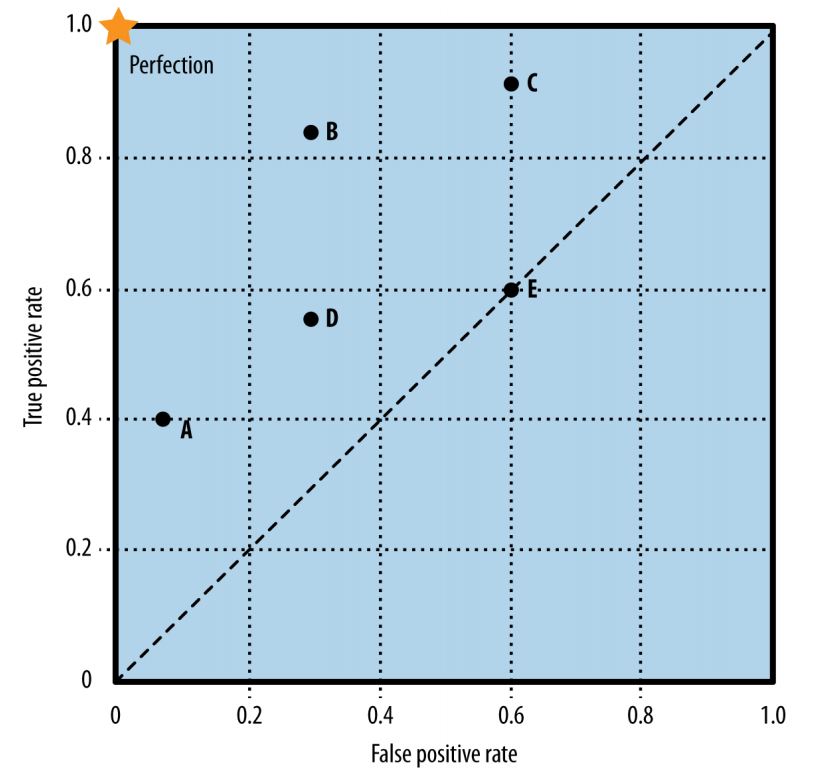
All methods except Gradient Boosted trees could not handle missing values. Preliminary runs of each method used subsets of the data without missing values. Then, imputing methods were trialled with gradient boosted trees since boosted trees can handle missing values. The MICE random forest method has been proven to work well with complex data sets (**???**). It works by first performing a standard random forest imputation of missing values which are treated as 'place holders'. Then, the imputations for one variable are deleted and the remaining full variables are used to impute the single variable using random forest imputation. This is repeated for each variable and their 'place-holder' imputations are replaced by an imputation targeted at the single variable. This cycle is repeated five times by default (**???**). If the imputed dataset gave similar predictive results using boosted trees as unimputed data using boosted trees, the imputed data must be reasonable. It is then trialled on the remaining methods and compared to un-imputed trials.

## Model Comparison, Model Averaging

To compare the models, mean squared error statistic was used for regression models and the area under the receiver operating characteristic (ROC) curve (AUC) statistic was used for binary classification. AUC is a better statistic than classification accuracy in comparing algorithms when the probability of the outcome is relevant to the problem. This is true for problems where the decision can depend on whether the probability is above a certain threshold (**???**).

An ROC graph plots the false positive rate on the x-axis and true positive rate on the y-axis where:

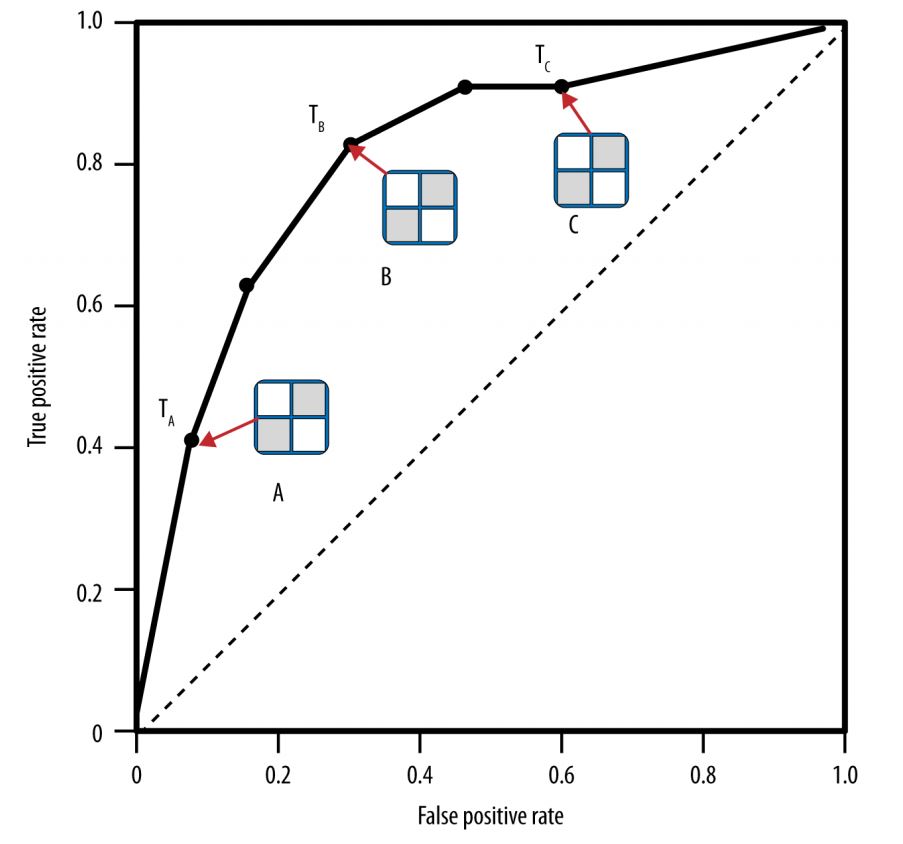
After a model has been created using a training set, the test set is run through the model, but in order to classify the test set, a probability threshold must be chosen. That is, the model outputs a probability to classify a case as 0 or 1. The probability is not a true statistical probability, but instead a measure of the certainty of the model. If a probability threshold is chosen to be 0.6, each case in the test set can then be classified as 0 or 1 depending on whether the probability calculated by the model is greater or less than 0.6. Once the classifications have been made, the true positive rate and false positive rate can be calculated for that model and threshold. This is plotted as a single point on the ROC curve, say point C. Refer below:



(Provost and Fawcett 2013)

A model that is perfectly classified would have all positives correctly classified (1.0 True Postiive Rate) and no incorrectly classified positives (0.0 False Positive Rate). If a model has a 1.0 TPR and 1.0 FPR (top right corner) this means it is correctly classifying all positives at the expense of incorrectly classifying all negatives as positive. It is 'dumbly' classifying all cases as positive. if a model classifies 0.8 of its positives as TP but also 0.8 of its negative cases as positive it is 'dumb' in a similar way to the previosu example. There is an 80% chance any case will classified positive. Therefore, models that lie on the diagonal line classify cases as well as random chance.

To plot the curve in an ROC, points are plotted for results on the training set using numerous threshold probability values between 0 and 1. Once enough points are plotted, a curve can be drawn. The closer the curve reaches the top left hand corner, the closer the algorithm is to perfectly predicting positive and negative cases. This would give an area under the curve (AUC) of 1. AUC of between 0.5 and 1 means the model is performing better than random chance for all possible probability thresholds.



(Provost and Fawcett 2013)

In order to compare whether models perform significantly better than one another, the AUC or mean squared error results from multiple-fold cross validations were used as the data points. Once the best two or three models were determined, these were combined using model averaging. Finally, the best model is determined between the single models and model averaged model.

## Bottom Line Translation

Although predictive power of the final model is important, for the method to be integrated into a company's decisions, the affect on a business' bottom line should be tested. This was done using a profit curve - a chart that plots the amount of 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 threshold were discounted. If the threshold is zero, all jobs are rejected and the profit is $0. If the threshold is 1.0, all jobs are accepted and the profit is 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 will mean result in higher profits, becasue jobs that are likely to make a loss are being rejected. This chart will clarify what percentage of profit increase the company can expect by integrating the algorithm into decision making.

A profit curve is made from a single train/test instance of the data. In order to understand the uncertainty around the profit curve, hundred-fold train/test splits were run and results calculated for each model and threshold point. This resulted in a statistical confidence interval around each threshold point which can be represented graphically. The final profit curve and surrounding confidence interval will indicate whether the algorithm makes a compelling case for integration into the company's decision making process.

# Results

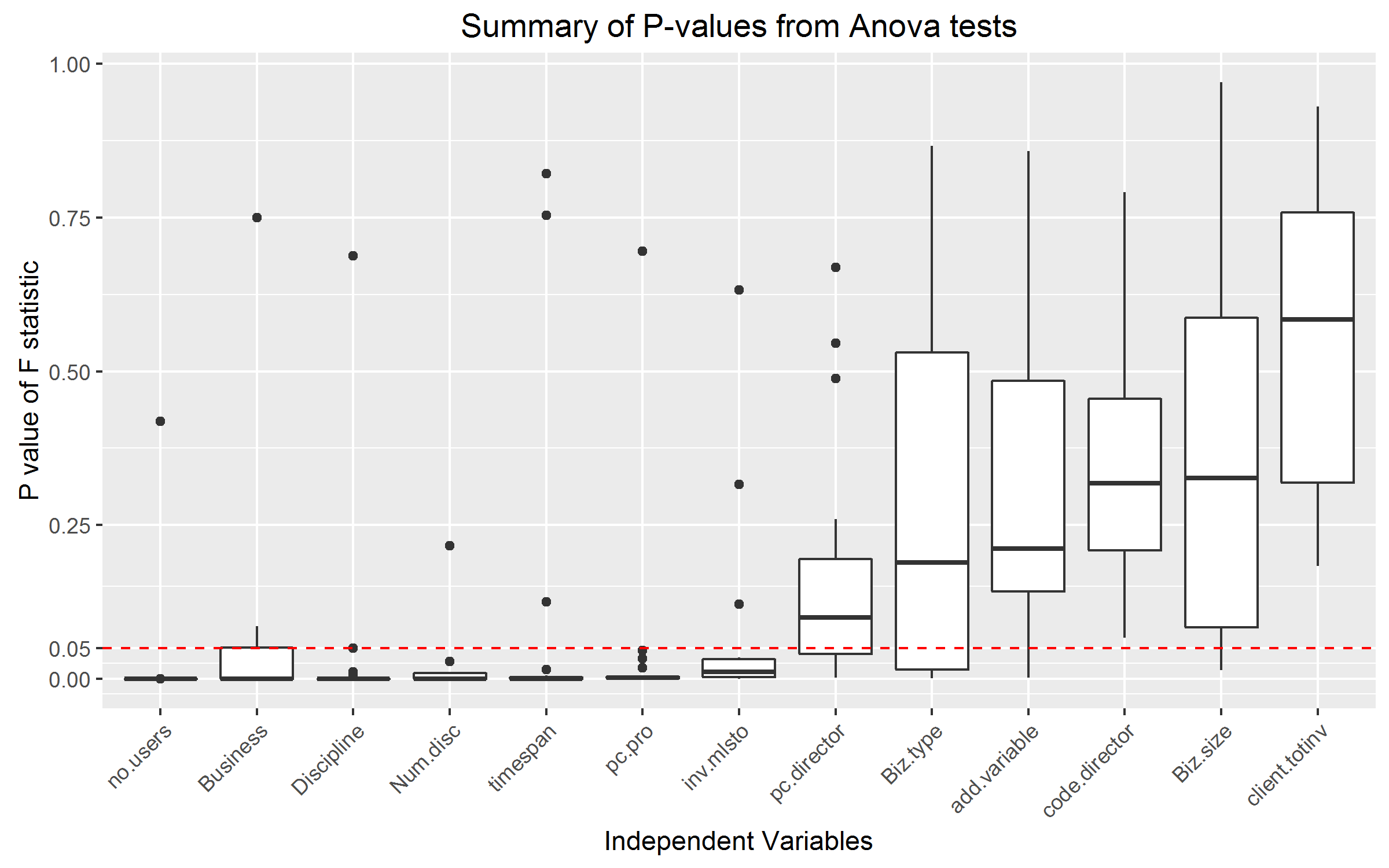
This section presents key charts and tables of results that progressively answer the research question: can a statistical or machine learning model based on historical project data predict the profitability of a new project and have a positive impact on the bottom line of the business?

## Variable Selection

To create the best possible model, only important variables were selected to reduce the affects of noisy independent variables. There are currently 46 independent variables for the single dependent variable, return per dollar which is the measure of profitability for a project.

### Regression

Regression requires complete datasets. Therefore, analysis was done with a core set of 13 complete variables and a single additional incomplete variable. That means a model was made for each incomplete variable, resulting in 21 separate ANOVA models. The p-values for the F statistic of each variable is plotted in the boxplot below.



The ANOVA model indicates the following variables have a median p-value below 0.05 and therefore significantly contribute to the rejection of the null hypothesis - the variables do not correlate linearly with the dependent variable, return per dollar.

* number of employees that contributed to the project (no.users)
* Percent of hours completed by a professional as opposed to technical employee (pc.pro)
* Business category of Client (Business)
* Discipline
* Number of internal disciplines involved in the project (Num.disc)
* Timespan of the project (timespan)
* Total amount invoiced for the project (inv.mlsto)

Four of the added incomplete variables (categorised as add.variable) were significant in their ANOVA models:

* Client id (code.client)
* Project category (JD.Second)
* Position of the main professional working on the project, ie mid-level, senior (ProjEng.Pos)
* id of main professional working on the project (code.ProjEng)

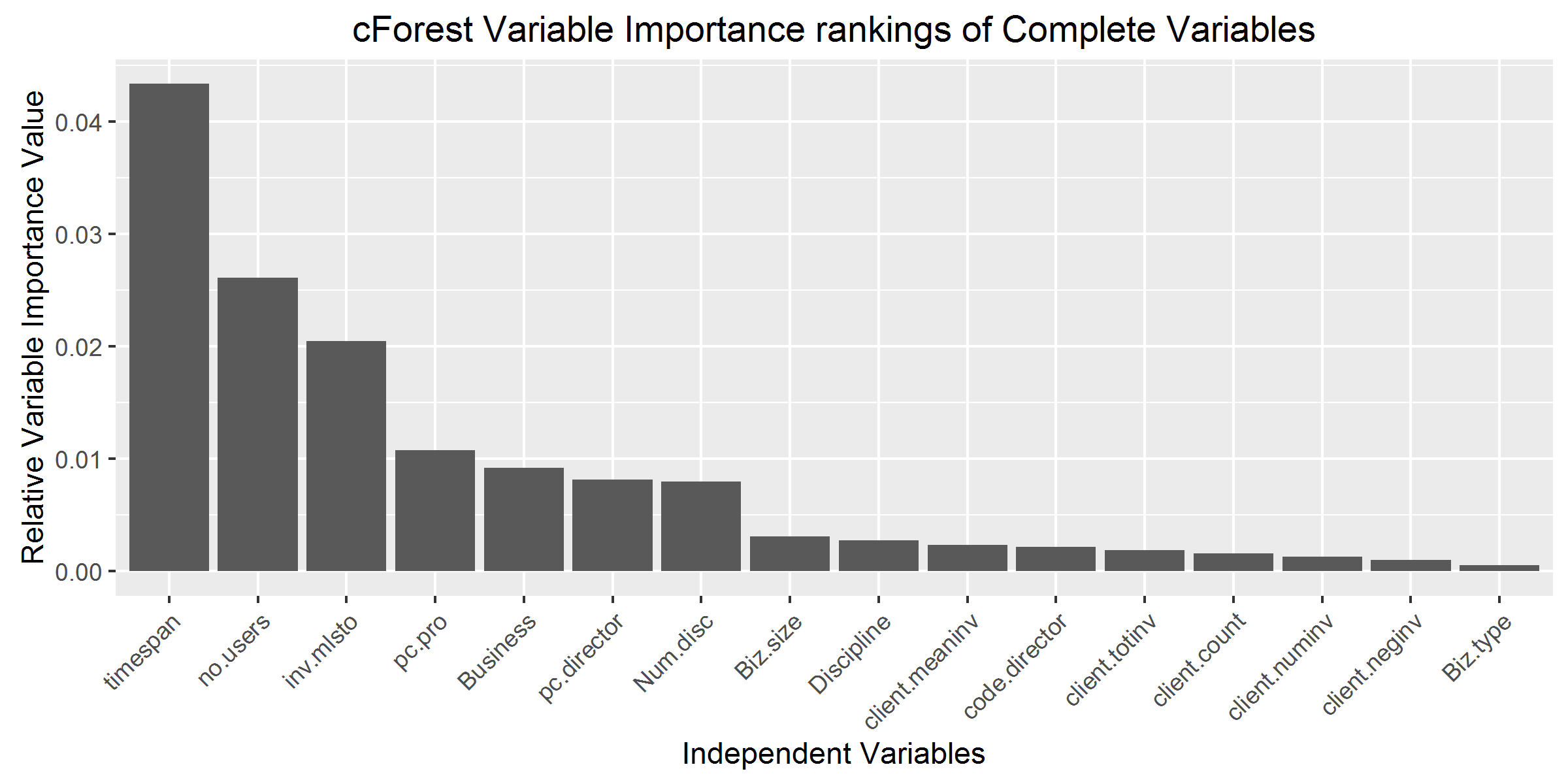
### Random Forest

The random forest algorithm imputes missing values, however this was not desirable for the early stage of variable selection - before modelling had commenced. Instead, a subset of complete data was used made up of variables with at least 2300 complete cases. Additionally, random forests cannot handle categorical variables with more than 53 categories which eliminated more variables. The resulting 'core' dataset contained 16 variables and 2364 cases. Because each run of the algoritm is slightly different, 10 algorithms were run, which each output the ranking of variables by importance. The mean ranking of each variable is summarised below:

|  |  |
| --- | --- |
| Variable\_Description | Rank |
| Amount Invoiced for Project | 1.0 |
| timespan | 2.0 |
| Number of employees on Project | 3.0 |
| Total Amount Invoiced from the Client - Past Jobs | 4.0 |
| % of Hours Completed by a Professional-level Employee | 5.4 |
| Business Category of Client | 5.6 |
| Mean Number of Invoices Submitted to Client - past Jobs | 7.2 |
| Mean Invoice Size for Client - Past Jobs | 7.8 |
| Number of Past Jobs with Client | 9.1 |
| Director | 10.4 |
| Number of Internal Disciplines Involved | 10.5 |
| Number Bad Debt Client Invoices - Past Jobs | 12.1 |
| Client Size in Number of Employees | 13.1 |
| Discipline | 13.8 |
| Broad Client Business Category (Gov, Private etc) | 15.0 |

### Cforest

Cforests were also run with a core of 13 complete variables, and incomplete variables were given added to the core variables individually in separate runs. Variables with an unlimited number of c ategories can be included in cforests which is an advantage over randomForests. Cforests compute a variable importance ranking which is a number relating to the reduction in error the variable provides. The variable importance rankings for the core 13 variables are as follows:



Some incomplete variables that were added one at a time ranked in the top 6 most important variables in their model:

* Percent of Hours completed by the main employee on the project (pc.majpos)
* Client id (code.client)
* Client contact id (code.contact)
* Project category (JD.Second)

### Variable Selection Summary

Based on the results of the three methods for variable importance, 11 independent variables were selected for the prediction modelling:

* Percent of hours completed by a professional level employee (pc.pro) -all models
* Project timespan (timespan) - all models
* Number of employees on Project (no.users) - all models
* Amount Invoiced for Project (inv.mlsto) - all models
* Business category of Client (Business) - moderately for all models
* Project category (JD.Second) - ANOVA and cForest
* Total Amount Invoiced from Past Client Jobs (clint.totinv) - randomForest
* Discipline - ANOVA
* position of main employee on the project (majority.pos) - newly created to cater for findings in ANOVA
* Percent of Hours completed by the main employee on the project (pc.majpos) - cForest
* Billing Type (hourly vs. fixed fee) - advice of Case Study partner

## Model Results

Once the independent variables were narrowed down to a succinct list, the prediction models were tested. First prediction return per dollar as a regression problem was trialled followed by prediction profit or loss as a binary classification problem.

### Regression Model

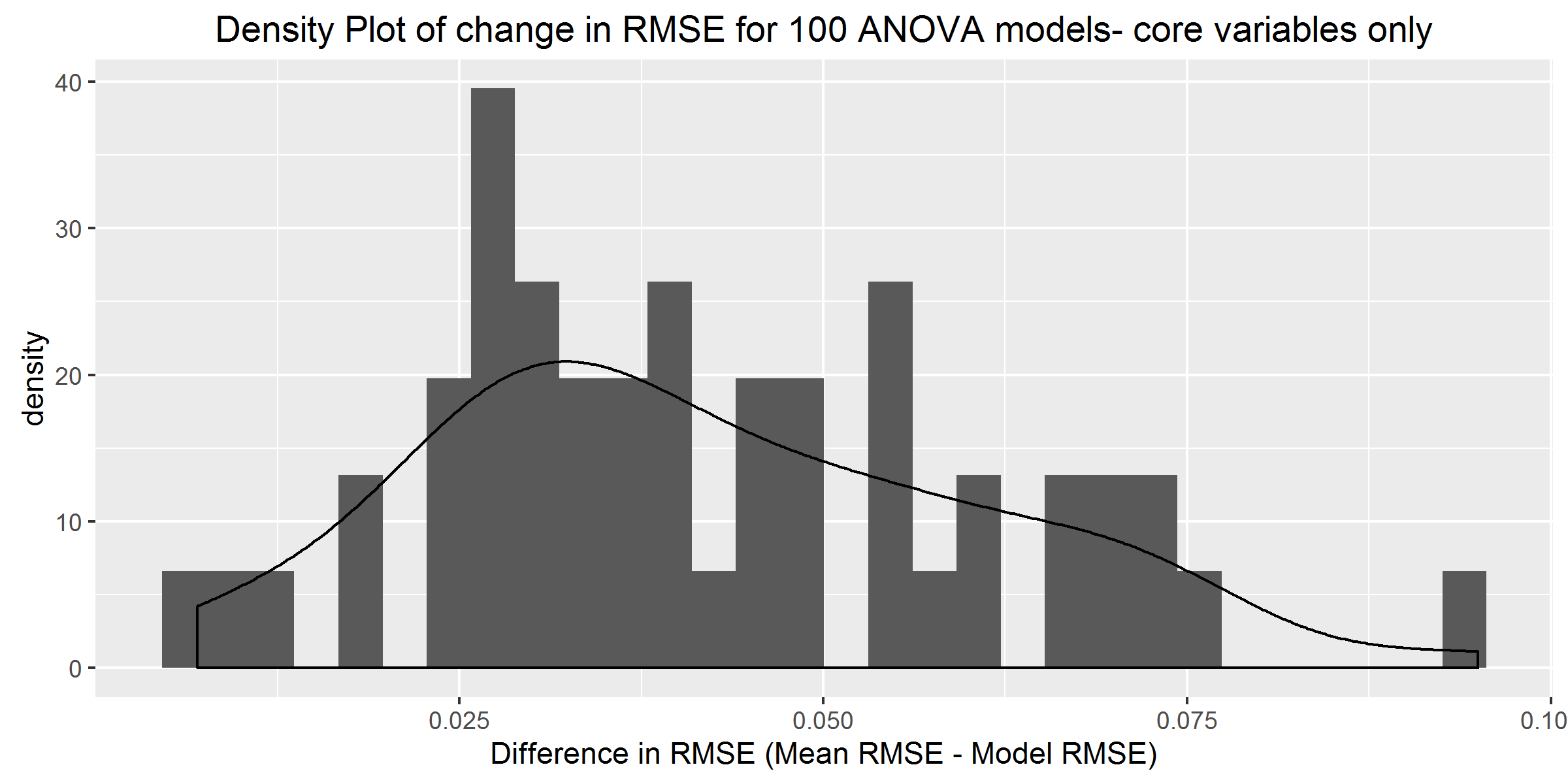
#### ANOVA

In ANOVA, linear relationships are tested and normal distributions are assumed for numeric variables. Therefore, where appropriate, numeric variables were normalised via log or cube root transformations. Trials of various variable interactions revealed that interactions between:

* total amount invoiced
* percent of job by professional level employee
* total amount invoiced
* client id
* project category
* number of employees on the job

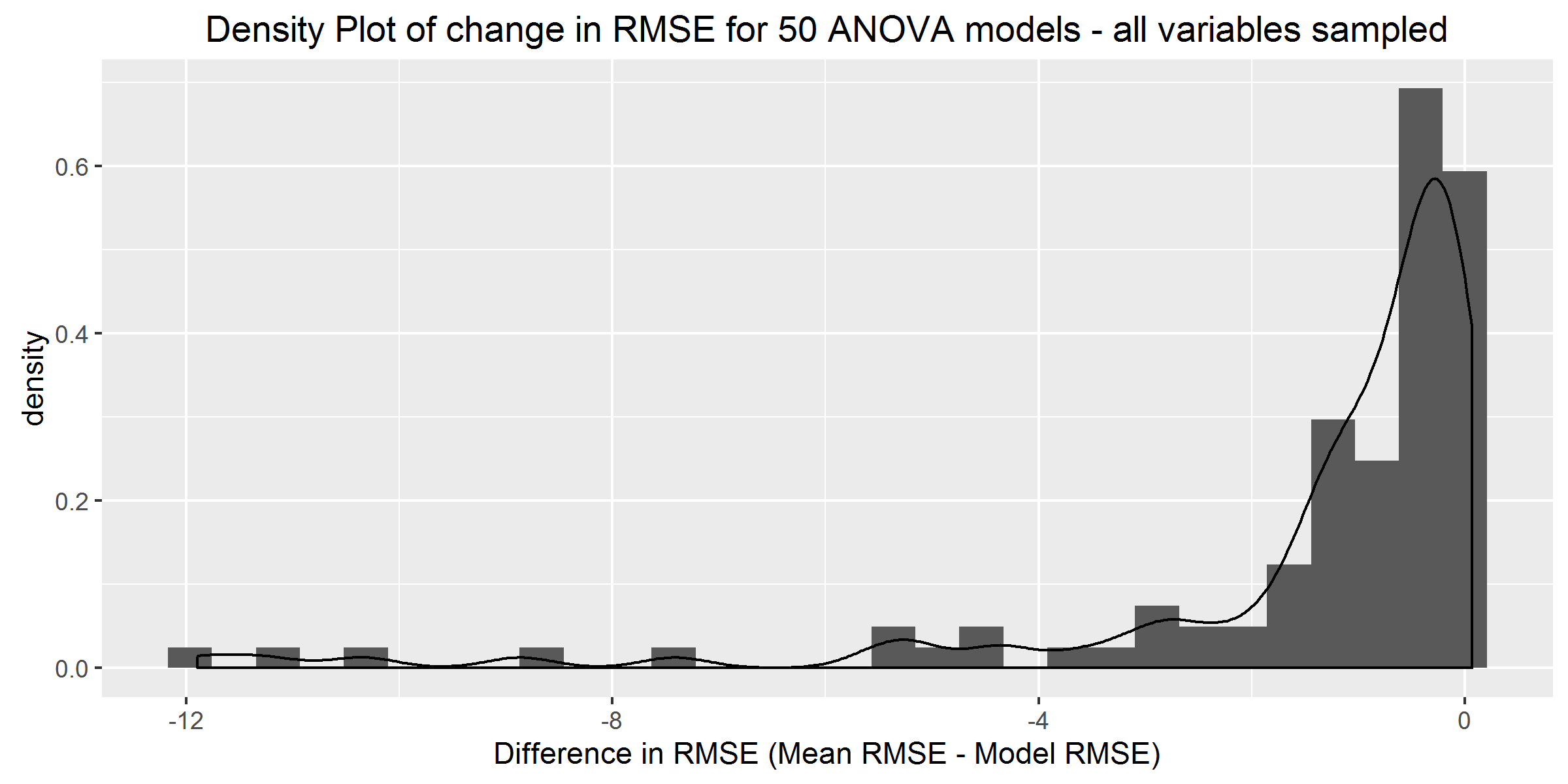
were statistically significant. Two types of ANOVA trials were run. The first used only a 'core' list of 6 variables that were largely complete and any relevant interaction terms. The second method was developed to get around the problem that when all variables are combined, only 15% of the projects have complete data. Therefore the ANOVA test would eliminate 85% of the data and imputing the data was not desirable at an early stage due to the amount of missing categorical data. The second method would choose a project (row in the dataset) randomly, and this case would have certain columns complete. The data was then filtered for all jobs that had at least the same columns complete, and the ANOVA test was run on this reduced set. This method allowed variables with a larger proportion of missing data to be analysed while not attempting to analyse all variables at once.

To assess the results of the ANOVA models, the root mean squared error (RMSE) of the return per dollar predictions were compared against the RMSE if all predictions were simply the mean of return per dollar for all projects. Then, the RMSE's from model predictions were subtracted from RMSE's using mean return per dollar. If the models were effective, RMSE should be lower for model predictions, and the difference would be greater than 0. Below is a histogram of this difference across 50 models run on randomly sampled 75% train and 25% test sets using core variables only. Note, only 18 samples (models) were required for a statistical power of 80% (Champely 2015).



A student t test indicates the difference is significantly above 0 with a p-value of 8.9e-21. Therefore, the null hypothesis that the difference is less than or equal to zero can be rejected. In other words, the ANOVA model improved estimates of return per dollar over using the mean return per dollar to a statistically significant degree. However, the mean difference is only $0.04. Furthermore, the mean error in return per dollar is $0.49 which is unacceptably high.

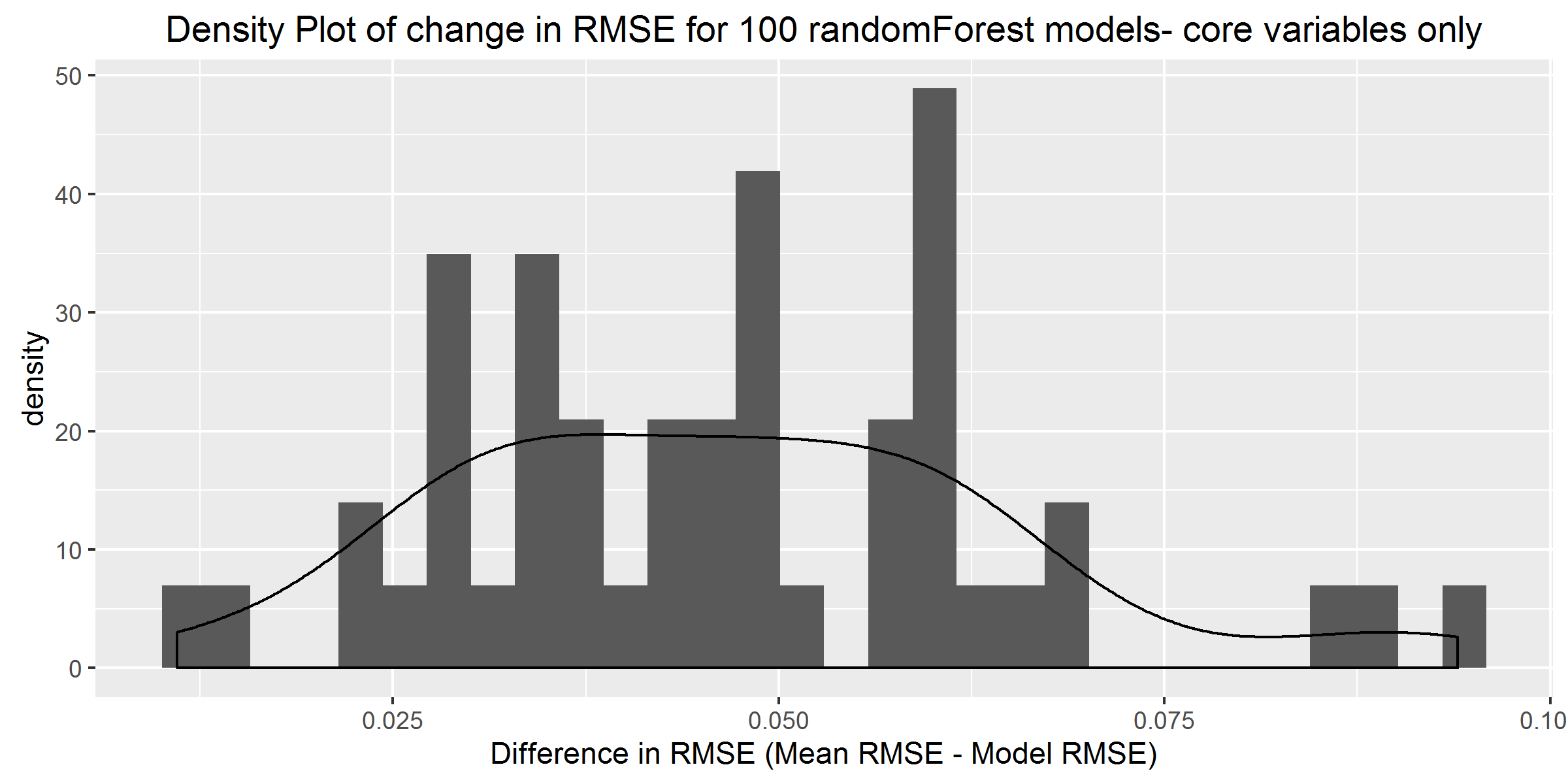
Results from 100 models using method 2, which captures incomplete variables, are shown below.



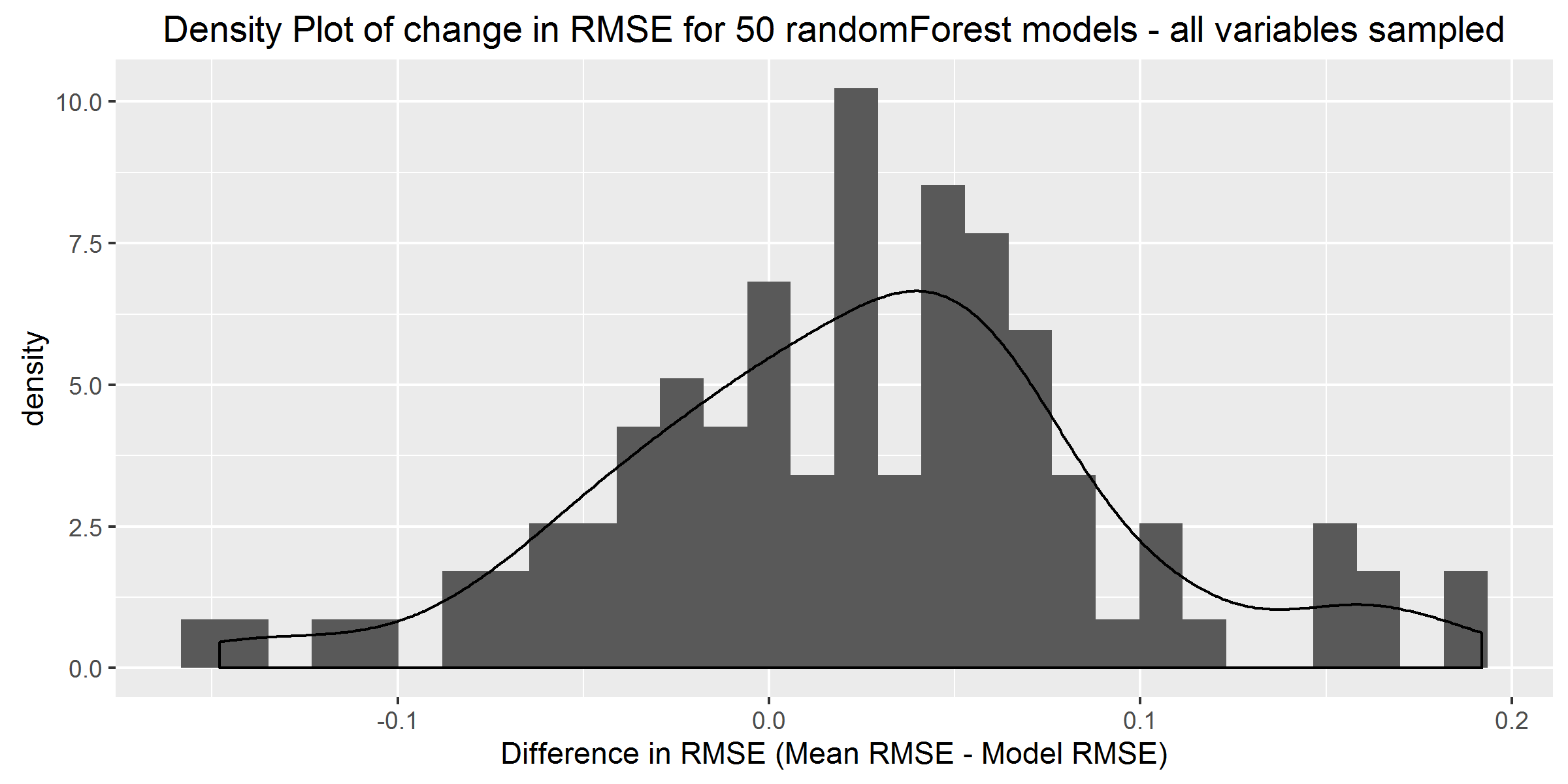
A student t test indicates the difference is not significantly above 0 which is clear from the plot (p-value = 1). This method clearly performed poorly in comparison to the core variables. A reason for this could be that the mean number of cases in the dataset was 2364 for the core variables but averaged only 876 for the sampled set of all variables. Regression results from ANOVA models were poor - randomForest was trialled next.

#### randomForest

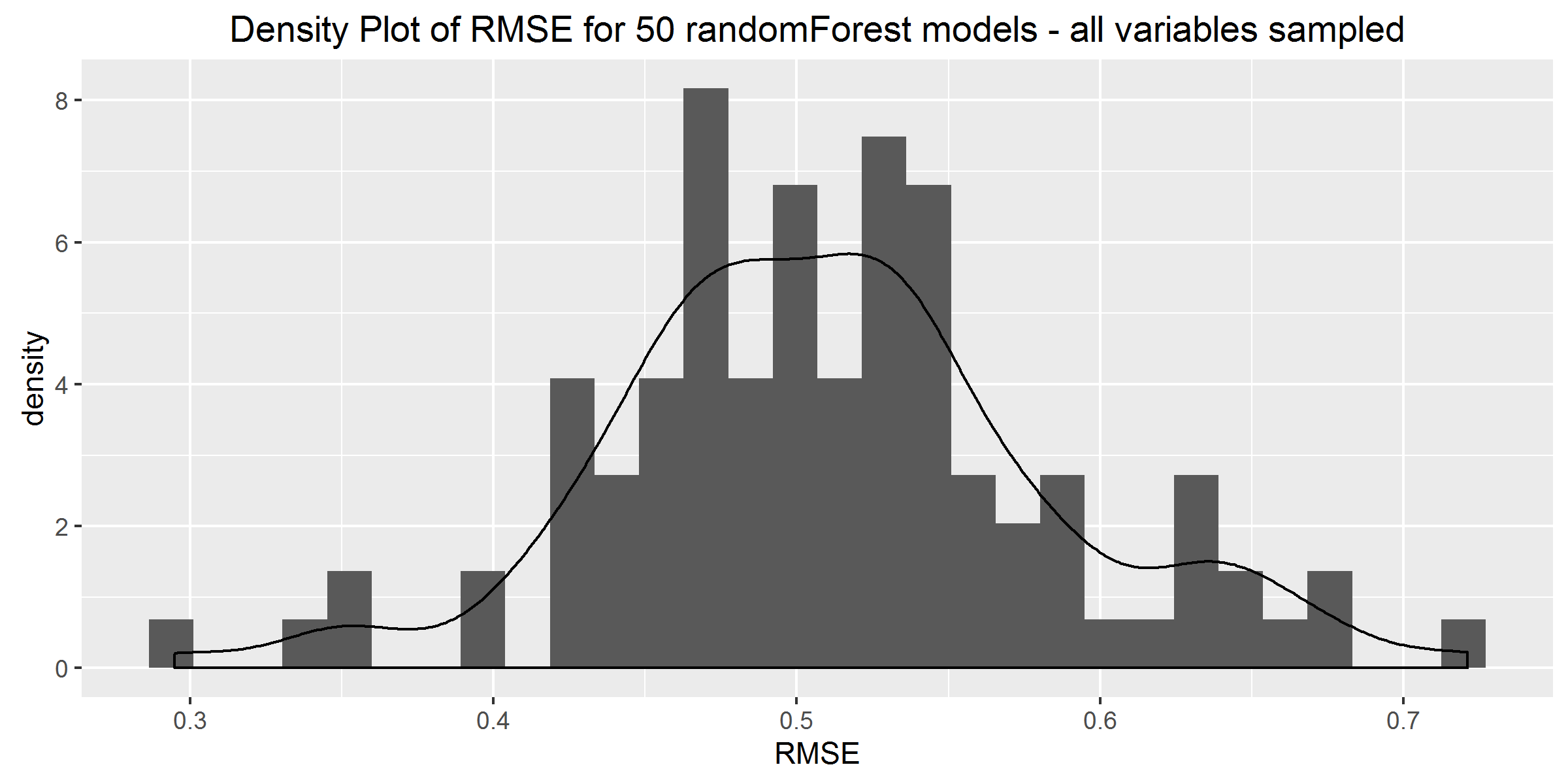
Numeric predictions of return per dollar were attempted first with core variables that were largely complete followed by a method of sampling a mix of core and less complete variables in every run. The randomForest parameters were first tuned using the caret package and found optimal values mtry = 5 and ntree = 500 (Jed Wing et al. 2015). A density plot of 50 models of the core variables using different train/test sets is shown below. The x-axis measures the decrease in RMSE obtained by using the randomForest model over simply the mean of all project return per dollar values. Note, only 45 samples (models) were required for a statistical power of 80% (Champely 2015).



This plot shows a mean return per dollar RMSE improvement of $0.04 which is the same as the ANOVA model. The p-value for the null hypothesis that the difference is less than or equal to zero is 6.4e-24 and is rejected. Results from 100 randomForest models using method 2 are shown below:



The p-value for these models and the afore-mentioned null hypothesis is 0.00013. The null hypothesis is rejected which contrasts the parallel ANOVA test. On the right hand side, of the graph above, the RMSE is improved by almost $0.20 but also worsens by over $0.10 per dollar. A density plot of the RMSE for the models is shown below:



Although the randomForest algorithm performed better than ANOVA when incorporating less complete variables, the spread of RMSE for return per dollar was wide and the mean was high ($0.51).

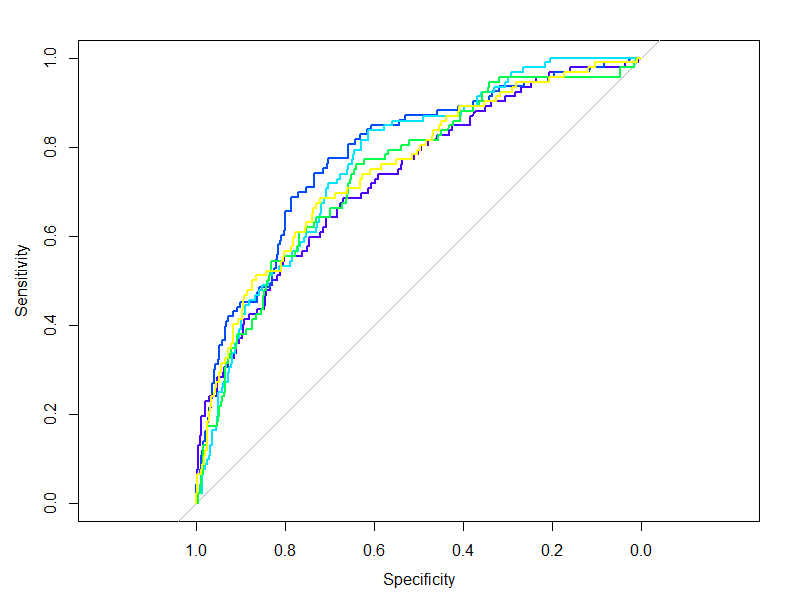
The two regression models predicting return per dollar resulted in high mean RMSE's ranging from $0.48 to $2. Because of the poor results, further development of regression models was postponed in favour of investiation of binary classification models. The classification is a simpler task than regression because it requires less precision.

### Classification Models

#### Initial Trial of five Methods

To begin analysis, all prediction methods were trialled using as much data as possible for each method. Boosted trees and Naiive bayes algorithm are able to process data with missing values, so all data could be used. Logistic regression, randomForest and Bayesian Networks require complete datasets so a subset of complete data was used for these. It was expected that Boosted trees would perform best as they are powerful and can use incomplete data. For this reason, 4 x 5 fold runs of each method were performed (20 models total for each method) and statistically compared against boosted trees. A power calculation was used to calculate the number of runs required for a statistical power of 0.8.

The models were compared using the Area under the curve (AUC) test statistic calculated from the Reciever Operating Curve (ROC) as this gives a good indication of model performance under binary classification (Robin et al. 2011)(Putler and Krider 2012). An example plot of five ROC's resulting from five fold cross validation models using the boosted tree algorithm is shown below:

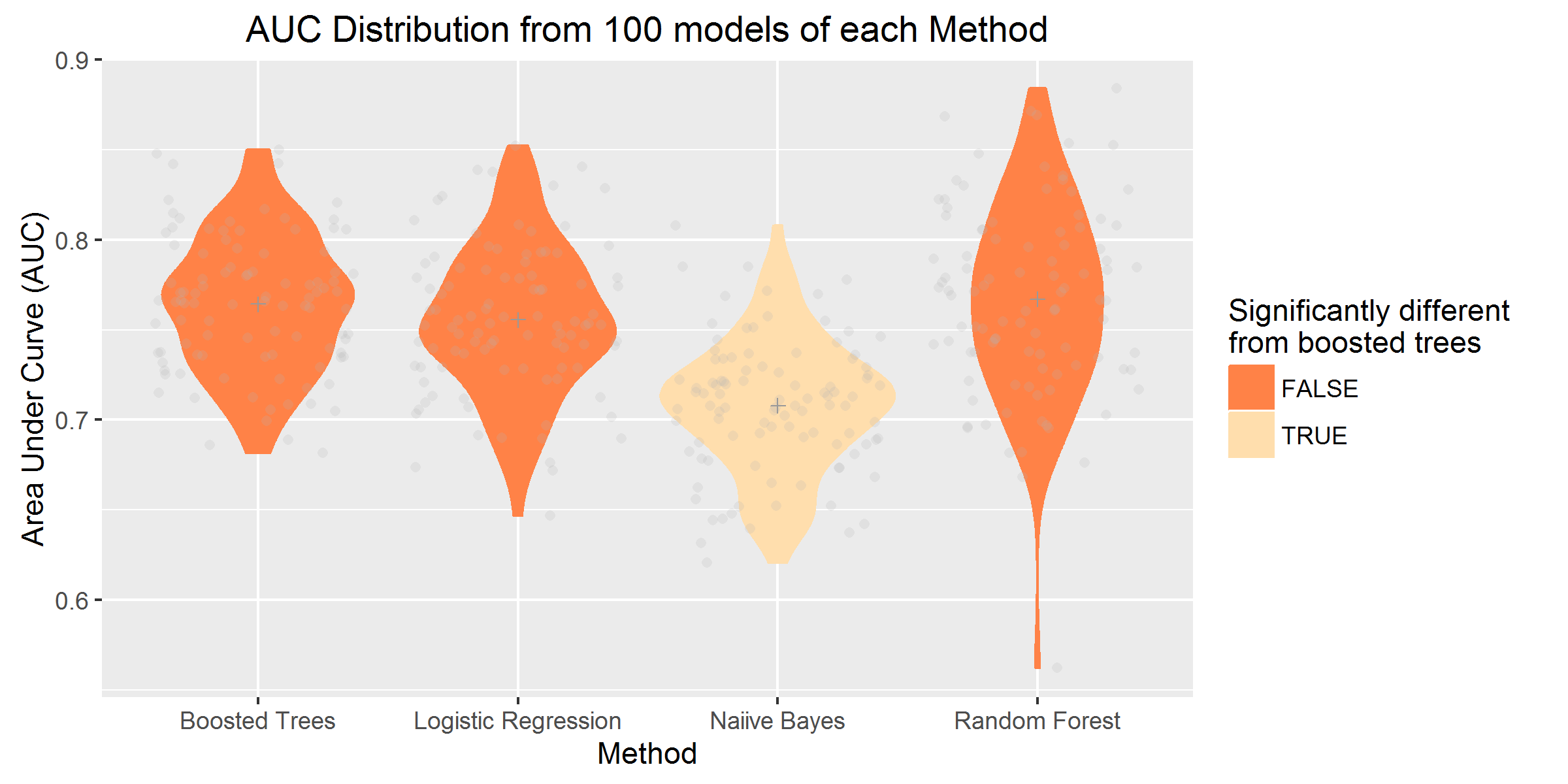


(Robin et al. 2011)

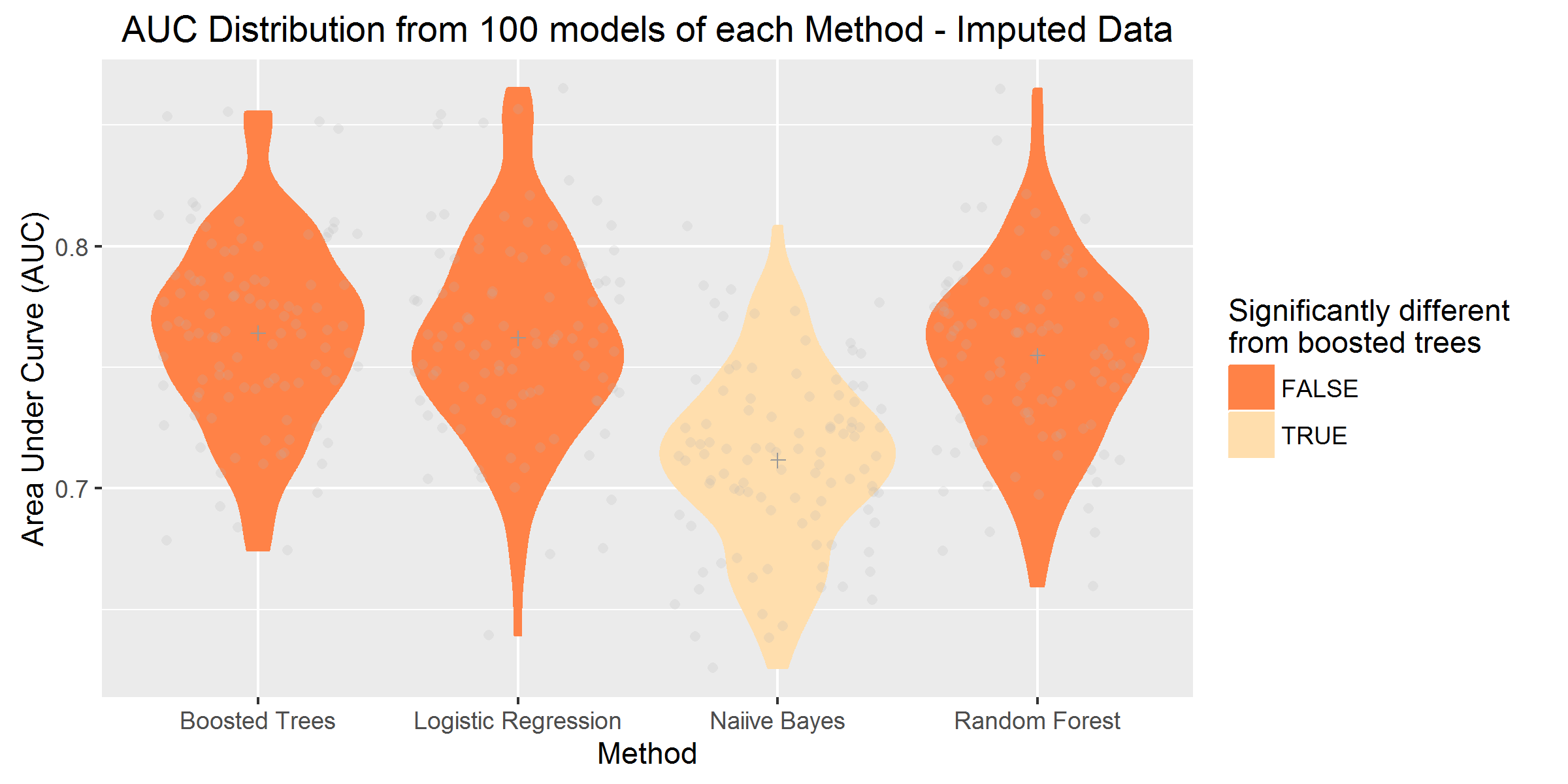
* Boosted trees first tuned using caret package (Jed Wing et al. 2015):
* shrinkage = 0.001
* no trees = 10000
* min terminal node size = 20
* max tree depth = 5

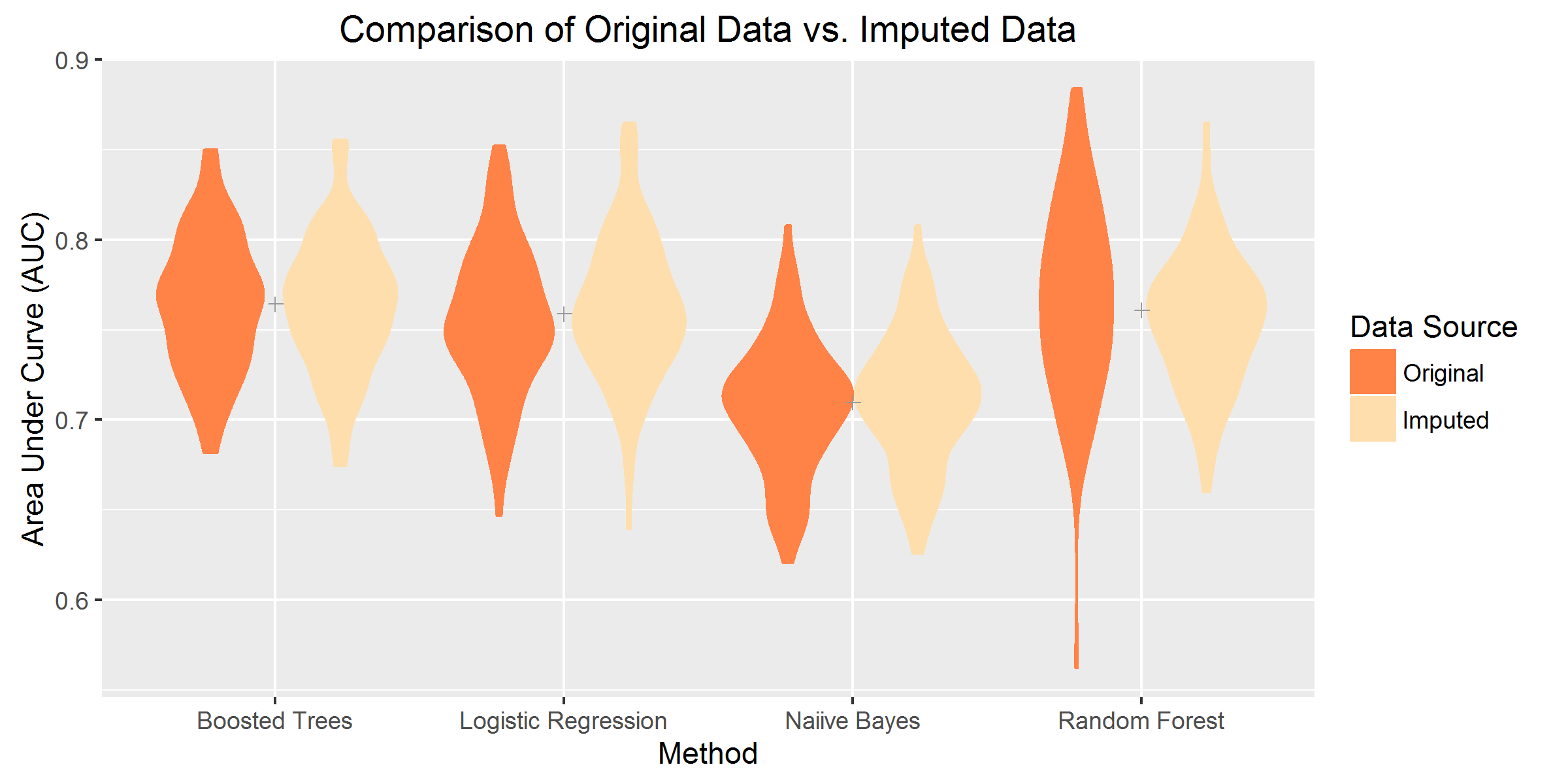
In this example, the average area under the curve (AUC) was 0.763 indicating that on the whole, taking into account all possible thresholds, the model is predicting profit/loss better than random chance (AUC = 0.5) but not perfectly (AUC = 1). 20 models and their resulting AUC statistics from each method were initially determined. These were used to calculate how many models were required to achieve a minimum statistical power of 0.8 when compared to the boosted tree results. The comparison of logistic regression required 100 models, while randomForest required over 700 because the AUC distributions were close. Bayesian Networks and Naiive Bayes did not perform as well and the power calculation revealed only 5 and 8 models were required to achieve a power of 0.8 respectively. Bayesian Networks were then dropped as Naiive Bayes perfomed slightly better and was fast and simple enough to work well as a baseline comparison to the higher performing methods.

It was finally decided that randomForest and boosted trees would likely not be distinguishable since the power calculations indicate 700 models are required. hence 100 models of each method were created as required for logistic regression. 10 fold cross validation was performed 10 times. The violin plot below summarises the AUC values from each method. The 'violins' are coloured according to whether the distributions significantly vary to boosted trees using a critical value of 0.05.



The performance of logistic regression and random forest algorithms cannot be statistically differntiated from boosted trees. The same process was repeated using an imputed dataset. Imputation was done using the MICE package and allowed all algorithms to be trialled with a full dataset. Again, 100 models were required to achieve a power of 0.8 although this could not be achieved for random forest unless over 450,000 models were made.





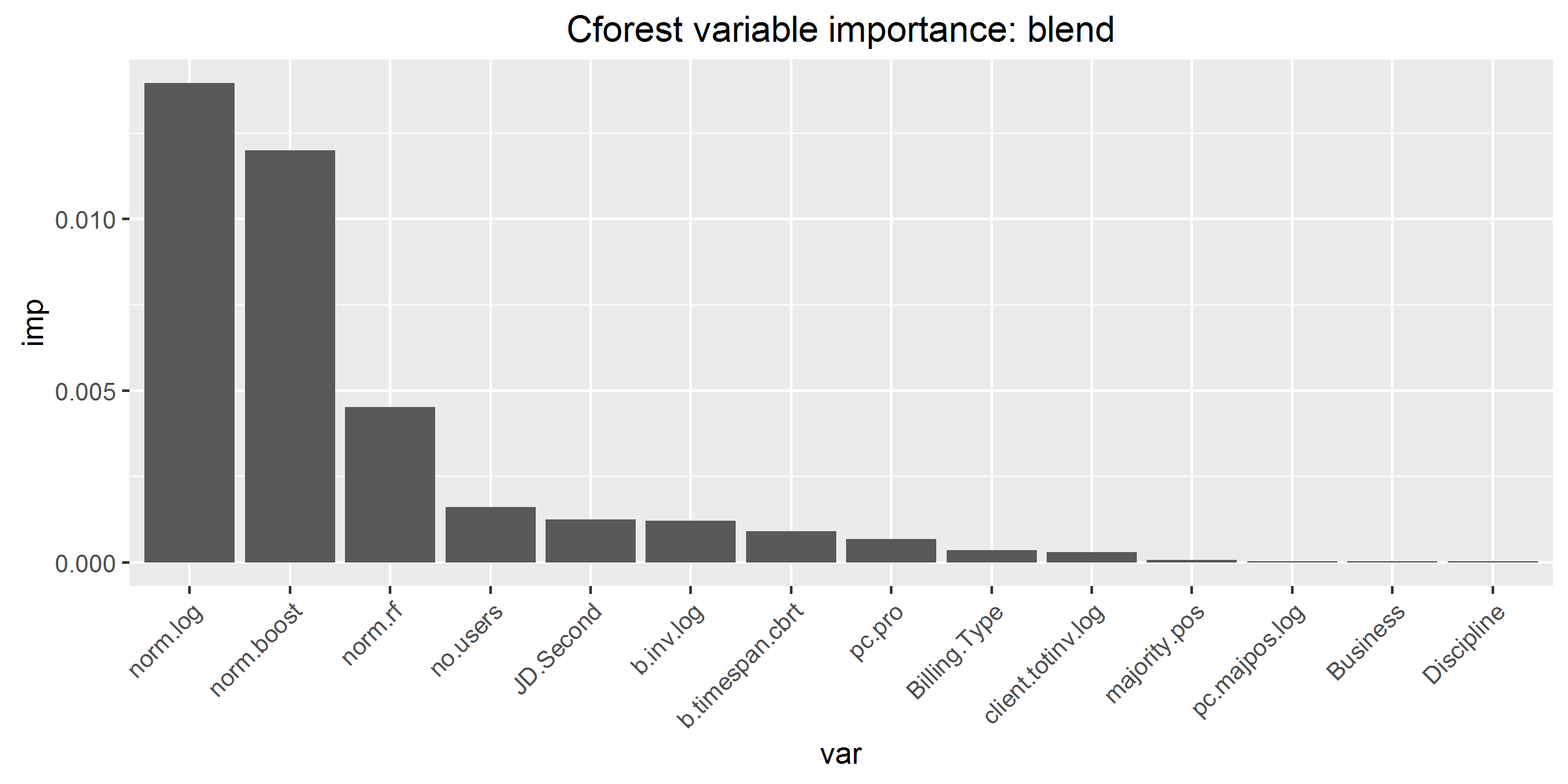
Imputed data tightens the distribution of randomForest substantially but slightly spreads the distribution of Boosted trees and logistic regression. Since the methods using imputed data have a similar performance, slightly tighter distributions of results and use more data, models using imputed data was used when developing blended models of all methods.

#### Blended Models

Blended models combine the results of each well-performing individual model to create averaged or 'blended' predictions. In this case, the randomForest, boosted trees, and logistic regression models were chosen for blending. The simplest way of combing models is to average the probability results (a number between 0 and 1) of all three models. Feature Weight Linear Stacking (FWLS), randomForest and boosted trees were all tested as blending methods where the predictions from each of the three single models becomes an additional variable in a new model. Each model blending method has its own method of weighting the outcomes from the individual models as a function of the values in the original model. For example, for a certain discipline in the case study company, randomForests may predict the outcome best so this is weighted higher for those cases. For this reason, a variable importance study was conducted again to simplify the models.

##### Variable Selection: Blended Model

*Variable Importance: cforest*



*Variable Importance: Logistic Regression*

Most significant variables (P value for coefficient < 0.05) are:

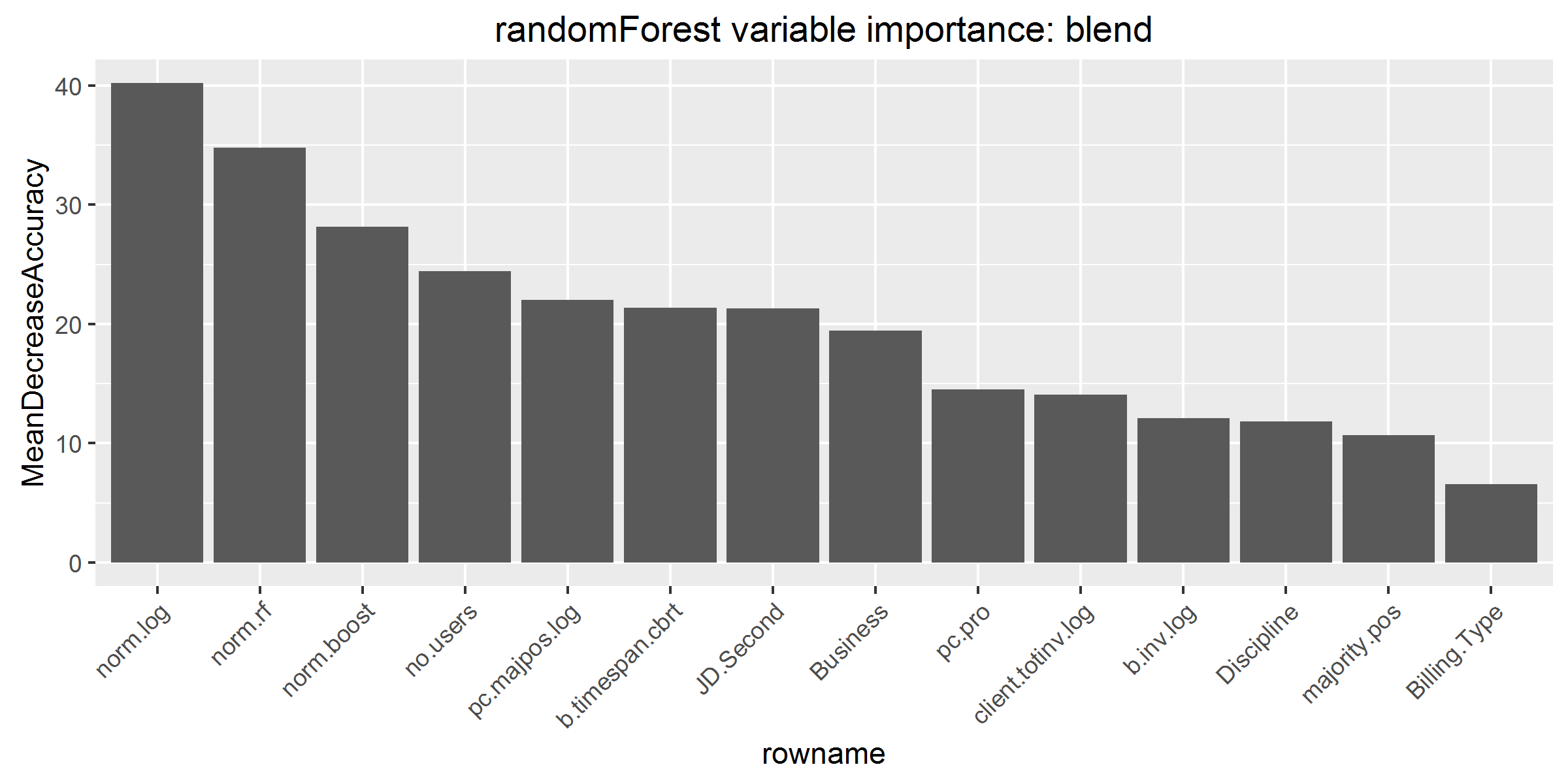
Interaction term between:

* Discipline \* results of logistic regression model
* Discipline \* results of boosted tree model
* number of employees on a job \* results of logistic regression model
* Client Business type \* results of boosted tree model

Single variables that aren't in interaction terms:

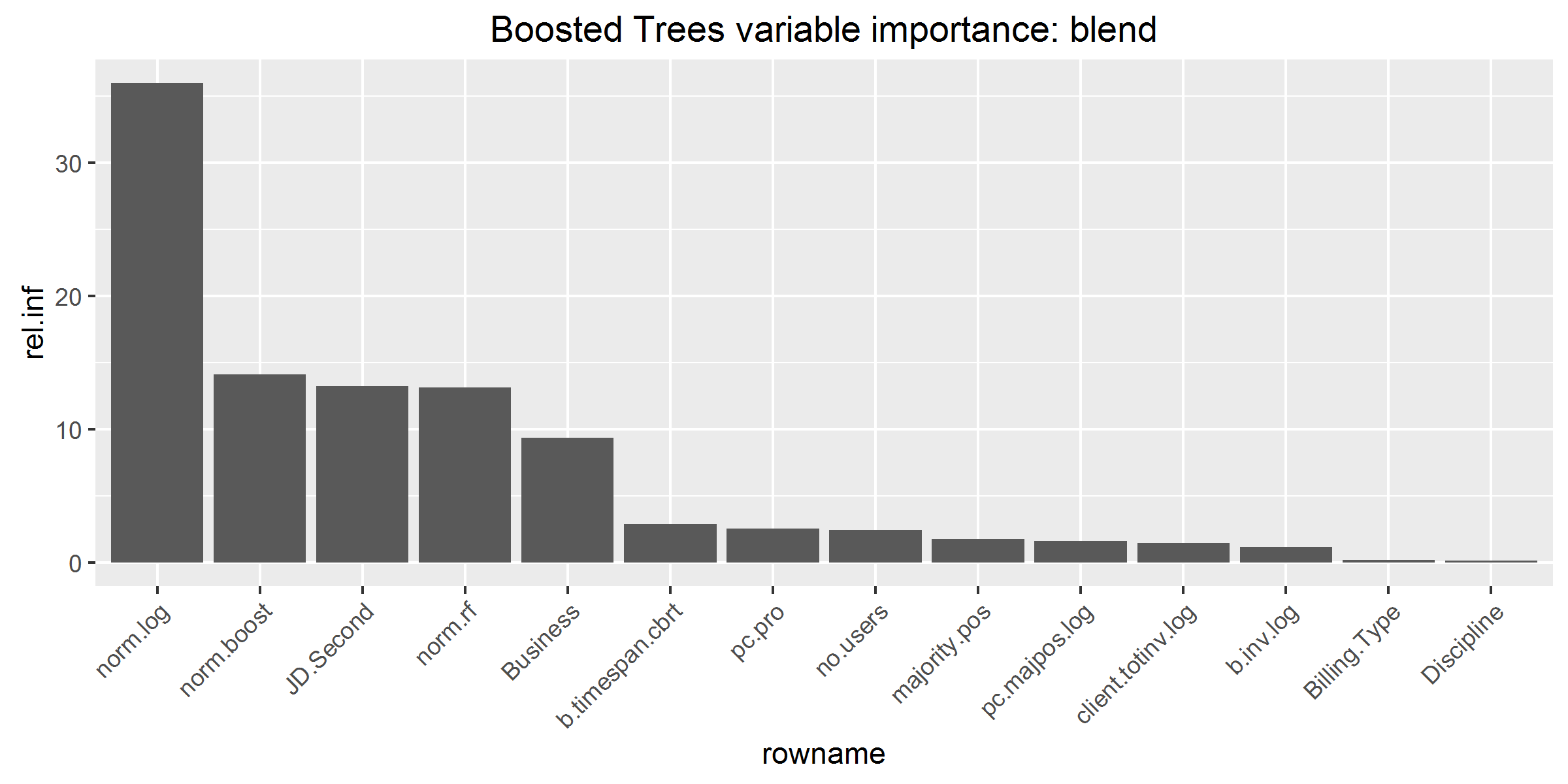
* Total amount invoiced for the job
* Timespan
* Project category

*Variable Importance: randomForest*



Permutation importance by randomForest rates all three prediction models as most important followed by number of employees on a project, percent of hours done by the chief employee on the project, timespan, and project category.

*Variable Importance: boosted trees*

 Boosted trees ranks all three individual prediction models highest followed by Project category, Client Business type, and timespan.

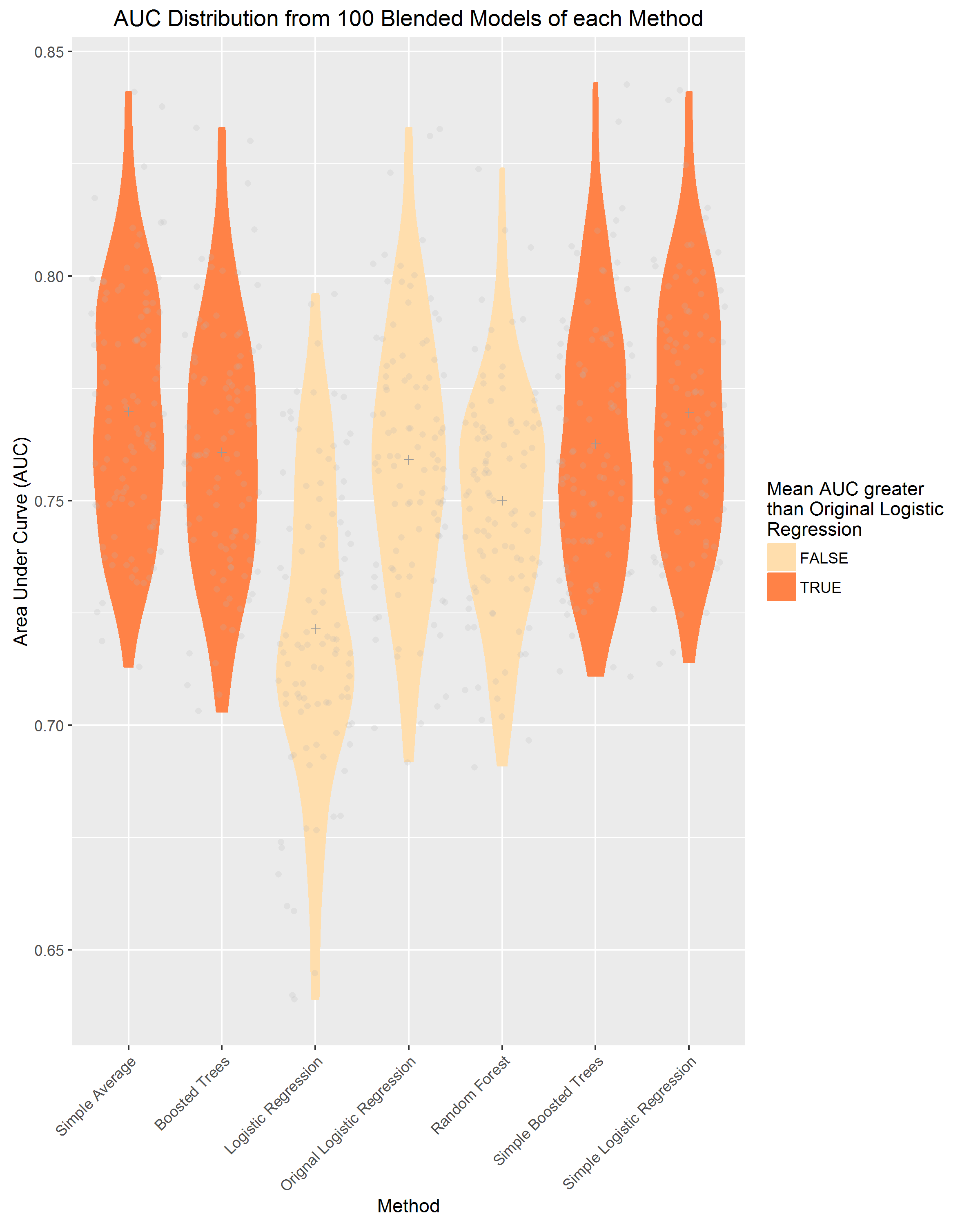
*Variable Importance Summary*

Upon reviewing important variables ranked by each method of blended models, the following variables were deemed important by at least two methods and shall be used in the comparison of blended models:

* all three results of prediction models, particularly logistic regression
* project category
* Client Business type
* Total Amount invoiced for a project
* Number of employees on a project
* Timespan

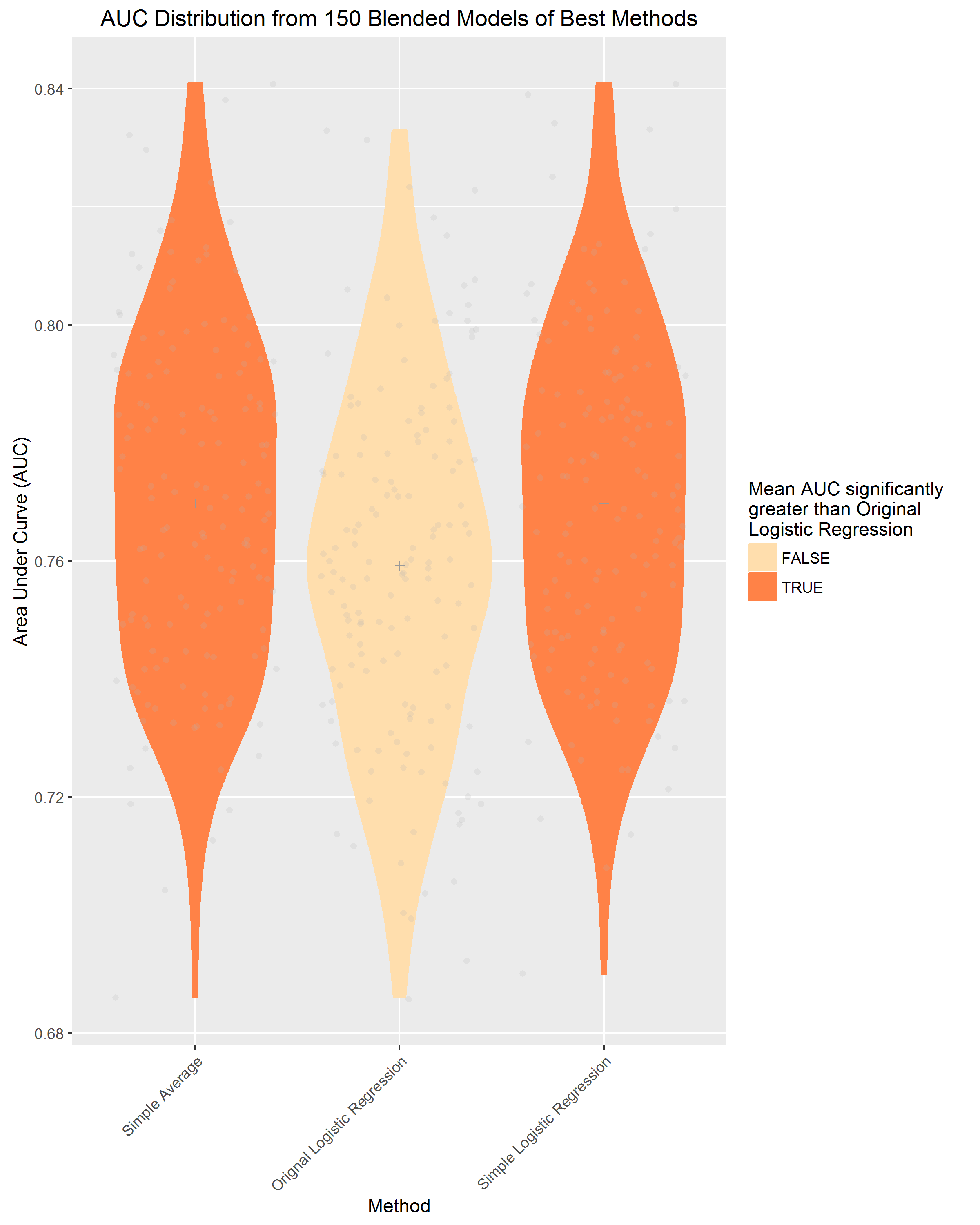
#### Comparison of Blended Models

Three model blending techniques were compared using the 8 variables determined above. These are FWLS, randomForest, and boosted trees. Furthermore, 3 simple model blending techniques were tested that used only the predictions of the three individual models as variables. These were averaging, a simple boosted tree, and a simple logistic regression. All six methods were compared against the original logistic regression predictions. Initially 100 train/test runs were performed for each method by running 20 iterations of 5-fold cross validation. The AUC results of these iterations are displayed below:



The plot shows that 4 methods could have a statistically significantly higher mean that the original logistic regression model: the simple blended logistic regression, the simple average, and the two boosted tree models. To achieve a power of 0.8, a two sample power analysis was performed on the 100 AUC results from each blended model against the AUC results from the original logistic regression. The boosted models required over 1000 samples while the simple averaged and simple logistic regression models required only 120. Therefore, 50 more iterations of the simple averaged and simple logistic models were performed since the boosted models results were not as promising.

The results of the 150 iterations are shown in the plot below. Simple model averaging results in a p-value of 0.0019 while the simple logisitc regression blended model had a two-sample t.test p-value of 0.0021.



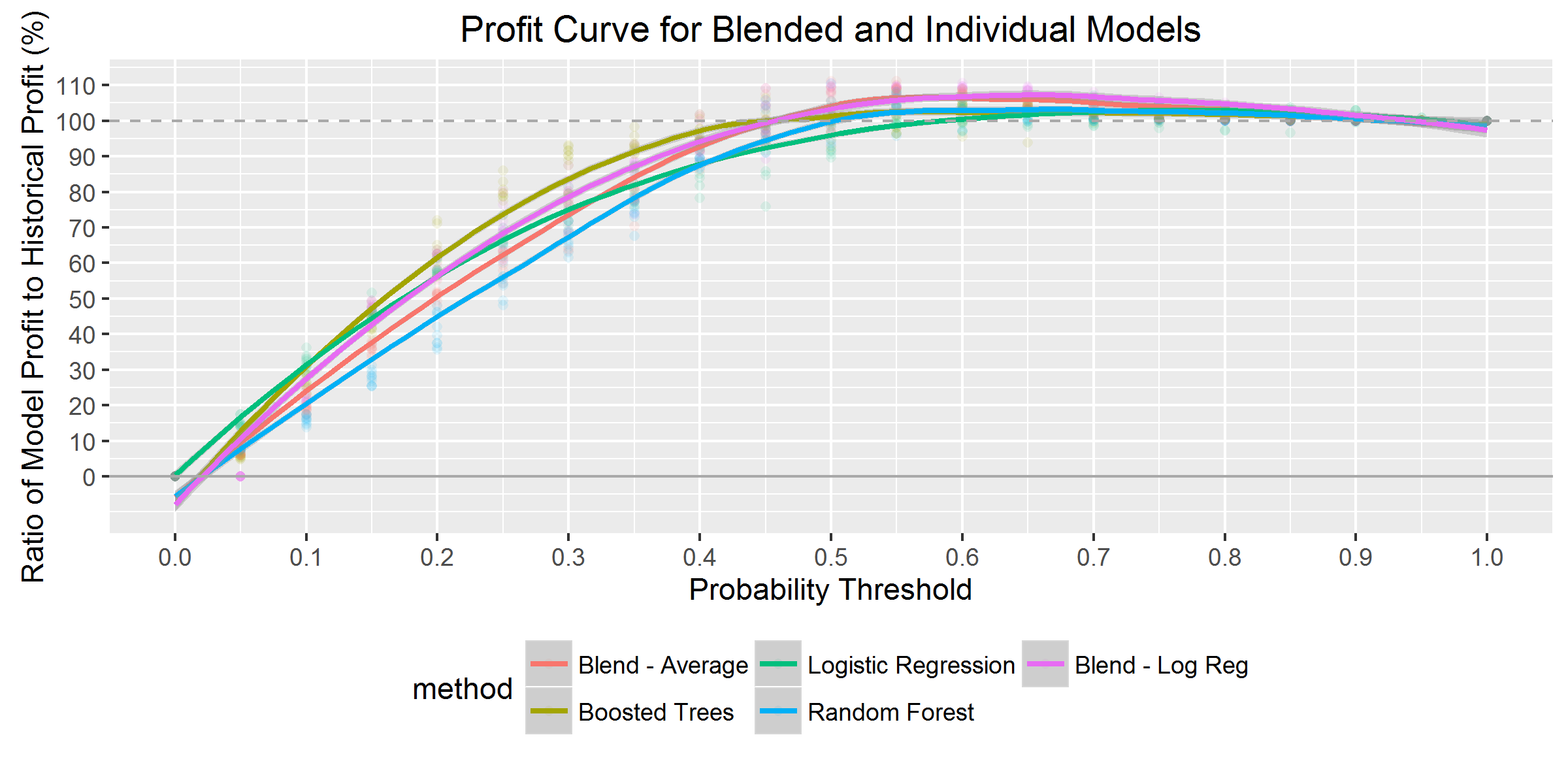
Simple logistic regression and averaging the the results of the three individual models (logistic regression, boosted trees and random forests) give virtually the same distribution of results. As averaging the results of the models is simpler than logistic regression, this was adopted as the best model.

### Bottom Line Analysis

The first research question, whether machine learning and statistical techniques can predict project profitability based on historic data, has been answered. The measure of profitability could not be modelled but simplifying the profitability prediction to a binary (profit/loss) model was successful. To answer the second research question, whether the model would improve the business' profitability, a bottom line analysis was performed.

The models output a 'probability' between 0 and 1 that each project will be a loss-making job (where probability = 1). This leaves the question, 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 rounding threshold, an experimental business-scenario was tested. At a given threshold, all projects calculated above the threshold were considered too risky, and were rejected. All profits and losses from these projects were forfeited, and the profits from the remaining jobs (below the threshold) were summed to give a total profit.

Using this logic, if the threshold was 0, all projects were rejected and total profit would equal 0. If the threshold was 1, all projects were accepted and the total profits would be the same as historical figures. Probability results from the two blended models as well as three individual models were tested and because the models require a training and testing set, the process was repeated 100 times with 20 x 5 fold cross validation models to achieve a statistical power of 0.8. The plot below illustrates the results of each model at each threshold with a local polynomial regression line of best fit (loess) and a 95% confidence interval around the mean.

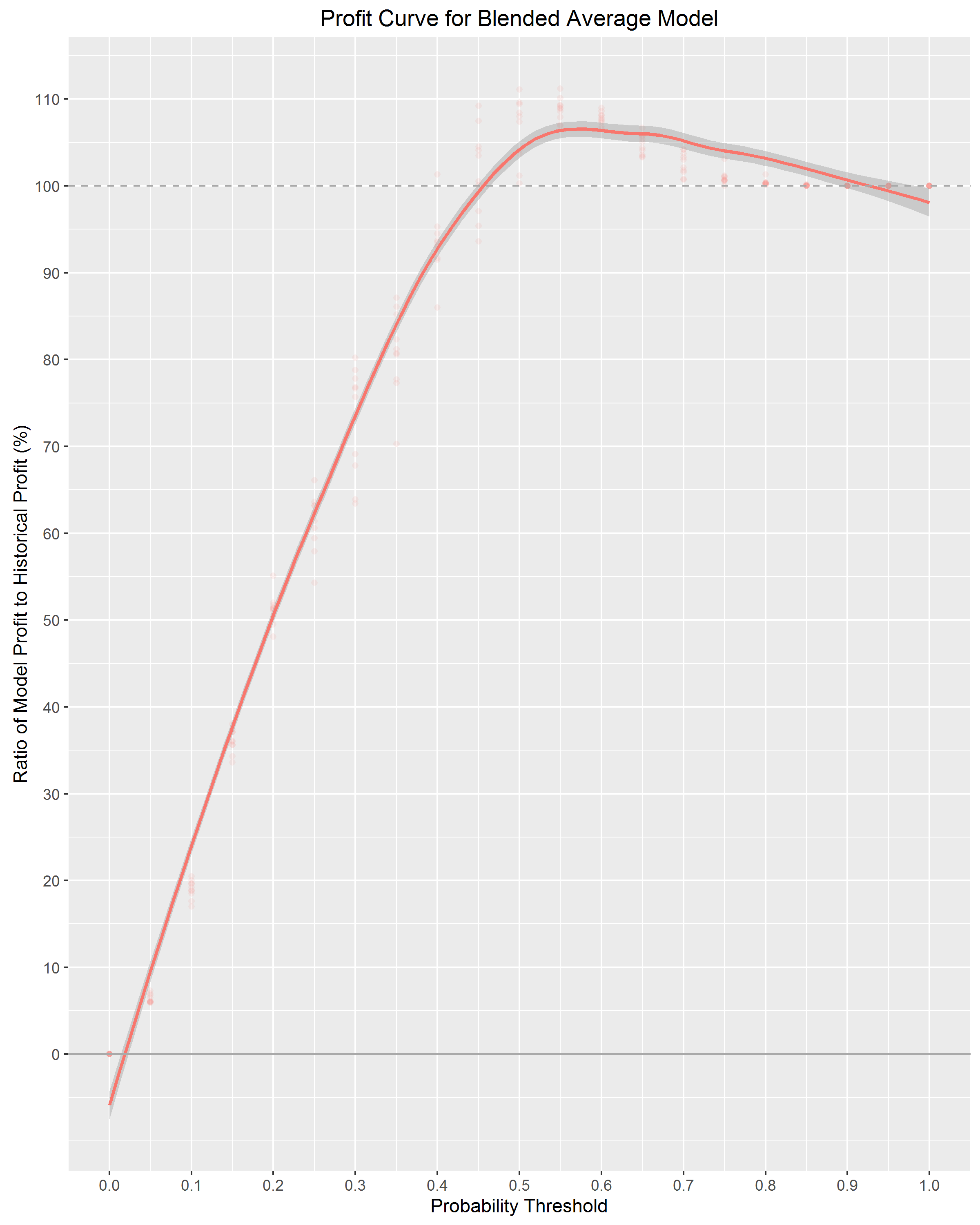


The two blended models outperformed the individual models as shown in this table describing each models' optimal thresholds:

|  |  |  |  |
| --- | --- | --- | --- |
| threshold | method | mean\_profit\_ratio | sd\_prof\_ratio |
| 0.55 | average | 108.8 | 1.3 |
| 0.60 | orig.boost | 103.4 | 3.7 |
| 0.75 | orig.log | 102.9 | 2.8 |
| 0.55 | orig.rf | 104.9 | 3.6 |
| 0.60 | simp.log | 107.6 | 3.0 |

The average of the three individual models performed best with the highest mean profit ratio at 108.8% and the narrowed standard deviation at 1.3%. This means that for the blended average model, if all jobs above the probability threshold 0.55 were rejected, the profits would increase on average by 8.8% in comparison to historical profit data.

For clarity, the plot for the average blended model is shown below:



# Discussion - 10 to 20 pages

## Variable Importance

The variable importance results from the three types of models: ANOVA regression, randomForest, and cForest were largely similar. All three models ranked five variables highly:

* percent of hours by professional
* project timespan
* number of employees in project team
* total amount invoiced
* business category of client which was positive

This was not necessarily expected particularly between ANOVA and the ensemble tree methods as ANOVA measures linear relationships as opposed to non-linear ensemble trees. Random forests favour variables that are numeric or have many categories and this held true for the analysis. The 6 highest ranked variables were numeric and one categorical variable, Business category of client, with 28 categories. The two lowest ranking variables had only 4 categories. It would be expected that cForests would rank the categorical variables with only 4 categories more fairly which was the case with Discipline. It was ranked mid way. Business category of the client and Broad business type however, were ranked similarly in randomforests.

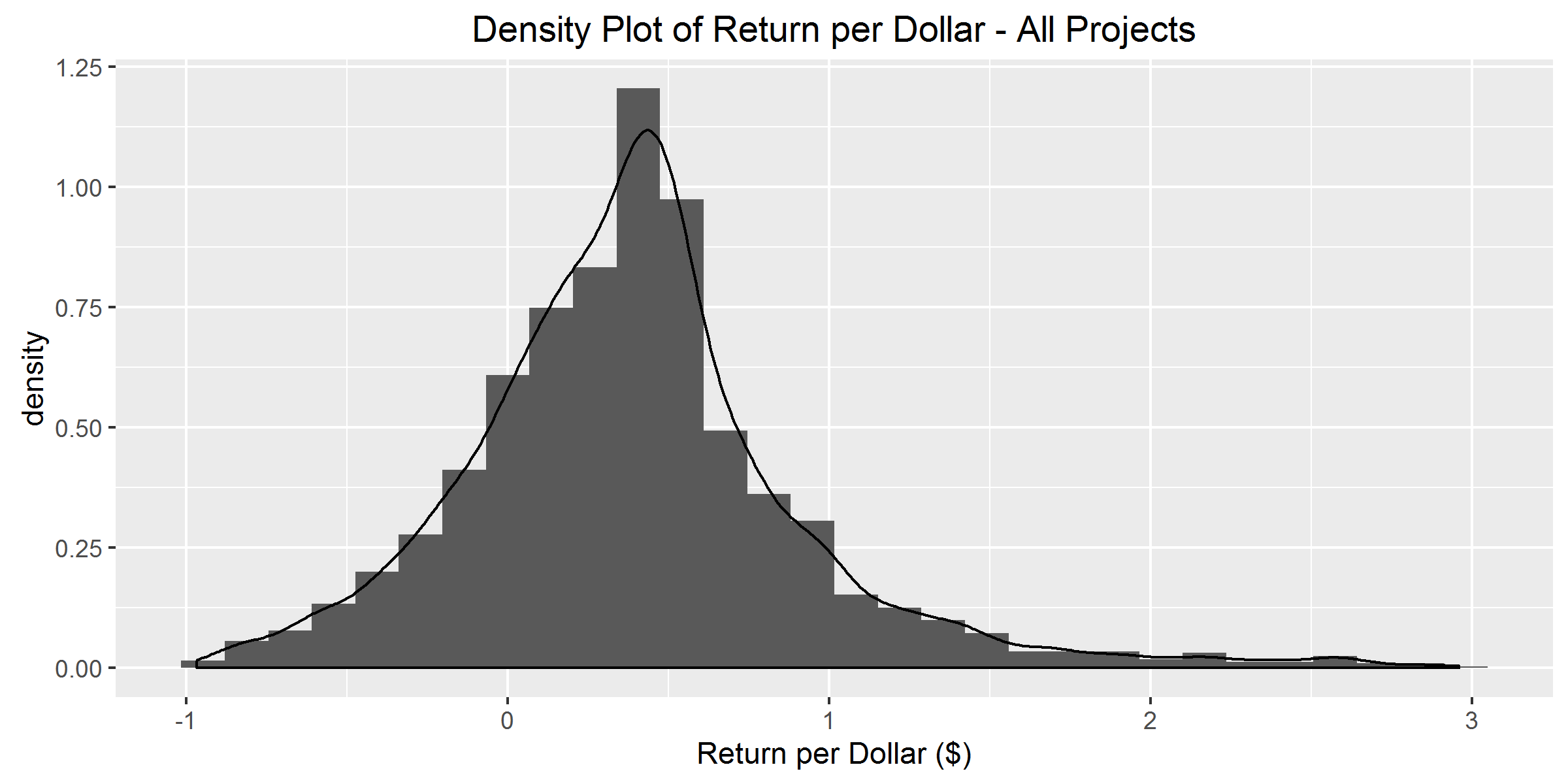
The nature of the higher ranked variables were not surprising, and were generally centered around who did how much of the work internally, general client characteristics and id, project timespan, and project category. It was expected that Billing Type would play a more important role in predicting project profitability as it describes the nature fees are charged. For this reason, and with advice from the case study business, the Billing Type was temporarily kept in the model.

For construction projects and IT software projects, the literature indicated size was the most important variable predicting effort. Size could mean size of a building or function point (level of functionality of the software) (Shepperd, Schofield, and Kitchenham, n.d.)(Finnie, Wittig, and Desharnais 1997)(Pai, McFall, and Subramanian 2013). In this case study, profitability is the dependent variable, not total effort and the size variable plays a different role. Instead of contributing to predicting the total effort, it is describing whether the size of a project correlates with how profitable the project is. The closest estimate of size in this case study is the invoiced amount which does show up as very important. The exact invoiced amount cannot be known before a project begins, as this would solve the problem of profitability. Therefore, the invoiced amount was binned into categories that a manager could choose as an estimate. The other important variables (including number of employees on a project, client business type and percent of hours by certain types of employees) have not been strongly represented in the literature to date.

## Regression

Both ANOVA and randomForest models predicted at best within an average of $0.48 for the dependent variable, return per dollar using either core variables or a sample of core and less complete variables. RandomForest predicts much better than ANOVA with a mix of core (almost complete) and incomplete variables while randomForest and ANOVA perform similarly with the core variables only. There is much less data available for training the models that include less complete variables so it is interesting that randomForest performed much better (RMSE of $0.50 for randomForest vs. $2 for ANOVA). The main difference between ensemble tree methods and ANOVA is that ANOVA assumes linear relationships between variables (some of which are normalised) whereas trees are capable of capturing non-linear relationships. It could be the case that the less complete variables had non-linear relationships with the dependent variable. RandomForest is also an ensemble method while ANOVA is not. Each randomForest is the average of 500 single trees, whereas each ANOVA model is simply a single model. 50 randomForest models produces 500 times more models than 50 ANOVA models so this averaging effect may be effectively eliminating outliers for better overall performance.

The RMSE can be evaluated in terms of the end-user's perspective (business manager who is proposing fees for new projects). There is a record of how previous projects have performed in terms of return per dollar. Refer below:



The 95% confidence interval for this data is $-0.65 to $1.43. If a project returns $0 for every dollar spent (return per dollar = $0), the project has earned only enough to cover the cost of doing the project (salary costs and business costs). If the return per dollar is -$0.50, the project has earned enough to cover half of the costs of doing the job. Finally, if the return per dollar is $1, the project has earned double what it cost to perform. If predictions have an average return per dollar error of $0.50, this represents an error in revenue of half what the project cost to complete. Clearly, this is not informative enough to bring into business decision making processes.

Further regression methods could have been persued, however randomForest is a high-performing algorithm and was far from providing usable results. It is likely the case that either not enough data was available or that the given variables do not capably explain the level of return per dollar. Each row of data is a project, and takes the company months, sometimes years, to create a single row of data. Therefore, more data cannot be accessed or easily created. It was decided categorical predictions should be pursued next as predicting whether a job will be profitable or not is a simpler task and could more effectively leverage the limited data.

Case studies in the literature tend to predict total cost/effort which is a regression problem (on a numeric scale) (Kim, An, and Kang 2004)(Chan and Park 2005)(Akintoye and Fitzgerald 2000). This is likely because the literature dealt with large projects which impact governments and large companies, and a simple profit or loss prediction is not useful. A regression prediction is not feasible for the data in this project, but a profit/loss binary prediction may be useful for this company because they complete hundreds of small projects a year. If binary classification is successful, the method the prediction contributes to company decisions must then be explored.

## Binary Classification

Testing gradient boosted trees, random forests, naiive bayes, logistic regression and bayesian networks in binary classification showed that boosted trees, logistic regression and random forests performed best (according to AUC). Bayesian Networks are not particularly popular and require discretised variables so it was expected that this algorithm would not perform as well as highly regarded methods such as boosted trees and random forests. Naiive 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 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 (**???**). This may explain why logistic regression is performing as well as enseble trees with the case study's smaller dataset with many variables.

Boosted trees, random forests and logistic regression all have AUC means around 0.775. The tests were repeated using imputed data and again the means sat around 0.775. The variables were not particularly well suited for imputing so this result is logical and could have even reduced the AUC. Many of the imputed variables were categorical and difficult to impute (such as job category, percent of job completed by a professional level employee, and the position of the majority worker on a job). The benefit of having an imputed dataset that performs well enough is that each model produces predictions for the full set of data. Beforehand, each model was limited to a complete set of data which implies a reduced data size. With complete model predictions, the high performing models can be blended to test whether better predictions are possible. Blending models brings about limitations which is discussed in the next section.

This individual models perform well above random chance (AUC = 0.5) and it can be expected that implementing a model would improve a manager's ability to predict whether a potential job will be profitable or loss-making. Whether the model is worth implementing in the work place is dependent on how the model affects decisions and the degree the algorithm will improve the 'bottom line' profits for the business. This will be reviewed in the following section.

## Blended Models

Trials of different methods of blending model results demonstrated that again, the simplest methods worked best with the data. In this case, 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).

It was plausible that simple blending models would perform better than complex methods because the data was not big enough for complex models to learn intricate patterns. This was observed in the single method models when logistic regression performed as well as random forests and boosted trees. The literature does not extensively cover blended machine learning methods however their success in the 2009 Netflix competition generated some publications. The Netflix dataset comprised of almost 3,000,000 observations and the size of the datacould have facilitated the success of complex blending methods such as FWLS (**???**). That being the case, it was still surprising that simple averaging of probabilities from the individual models performed as well as blending results using logistic regression. It was expected that logistic regression would have an advantage over model averaging due to its ability to weight the prediction variables while maintaining a simple model structure.

The logistic regression blending model did indeed weight predictions from the original models differently. If the predicted probability from the original logistic regression model increased by 0.1, the odds of the model prediciting a loss making job increased by 37.4%. Similarly when probability predictions from the random forests model increased by 0.1, the odds of the blended model predicting a loss increased by 25.8%. On the other hand, for the same increase in the original boosted tree model, the odds of the blended logistic regression model predicting a loss increased by only 2.6%. This contrasted with the alternative simple model that averaged the probabilities from each original model (i.e. weighted them equally) but still managed to produce similar AUC results. It was not clear why this was the case, however both simple models were carried forward to analyse their respective bottom line impacts on the company's profits which could reveal performance differences.

## Bottom Line Analysis

The blended models significantly (?) outperformed individual models which was expected given the two blended models' AUC results were significantly higher than the original logistic regression results (which were in line with random forests' and boosted trees' results). The simple average of logistic regression, random forests, and boosted trees results produced a better profit curve than a logistic regression of the same three variables. The model produced a 8.8% increase in profit by rejecting jobs with a probability of loss over 0.55. To put this in context, if the company's yearly profits were $1,000,000, the mean yearly increased profit would be $8.810^{4} by rejecting all jobs with a probability higher than 0.55. In the data, projects assigned a probability higher than 0.55 represent only 4.2% of all projects.

Alternative strategies exist for how to deal with projects that are marked with a probability of loss greater than 0.55. For example, contracts for these projects could be changed to hourly rate contracts without exceptions. This means that each hour booked on the project will definitely be charged to the client for each stage of the project. Another example would be to increase fixed fees for projects above the threshold by a nominal percentage. However, this method seems less practical as the case study projects above 0.55 would need to increase fees by 46% in order to generate a 15% profit. A proposed fee increase of that magnitude may not be as acceptable as an hourly rate contract depending on the project and client. Contracts could be adjusted in many other creative ways to reduce risk in high probability projects, and the threshold of 0.55 was optimised based on the accept/reject strategy. Different courses of actions could be developed for specific clients or project sizes via analysis of threshold scenarios.

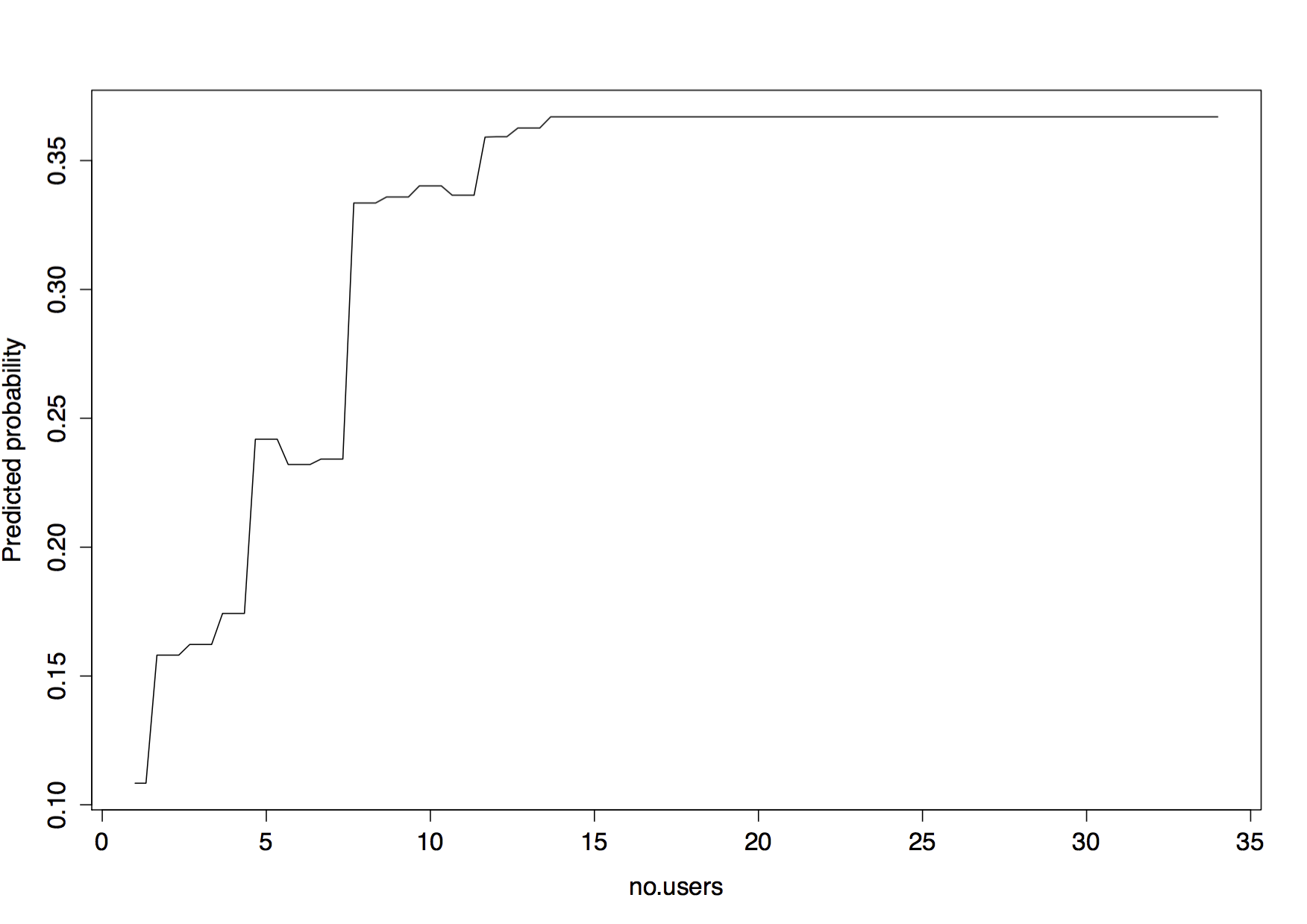
Because of the positive AUC results, it was expected the models would translate into financial benefits for the company. The results look reasonable, profit gains may be high enough to trigger further cost-benefit analyses for introducing the model into business decisons. It was surprising that, given the similar AUC results, simple model averaging signficantly (?) outperformed the blended logistic regression model's profit curve. This means that model averaging predicted projects with probabilities closer to 1 (loss-making jobs) with more accuracy than blended logistic regression. This may be because ... ROC curves.

### Limitations and Future Work

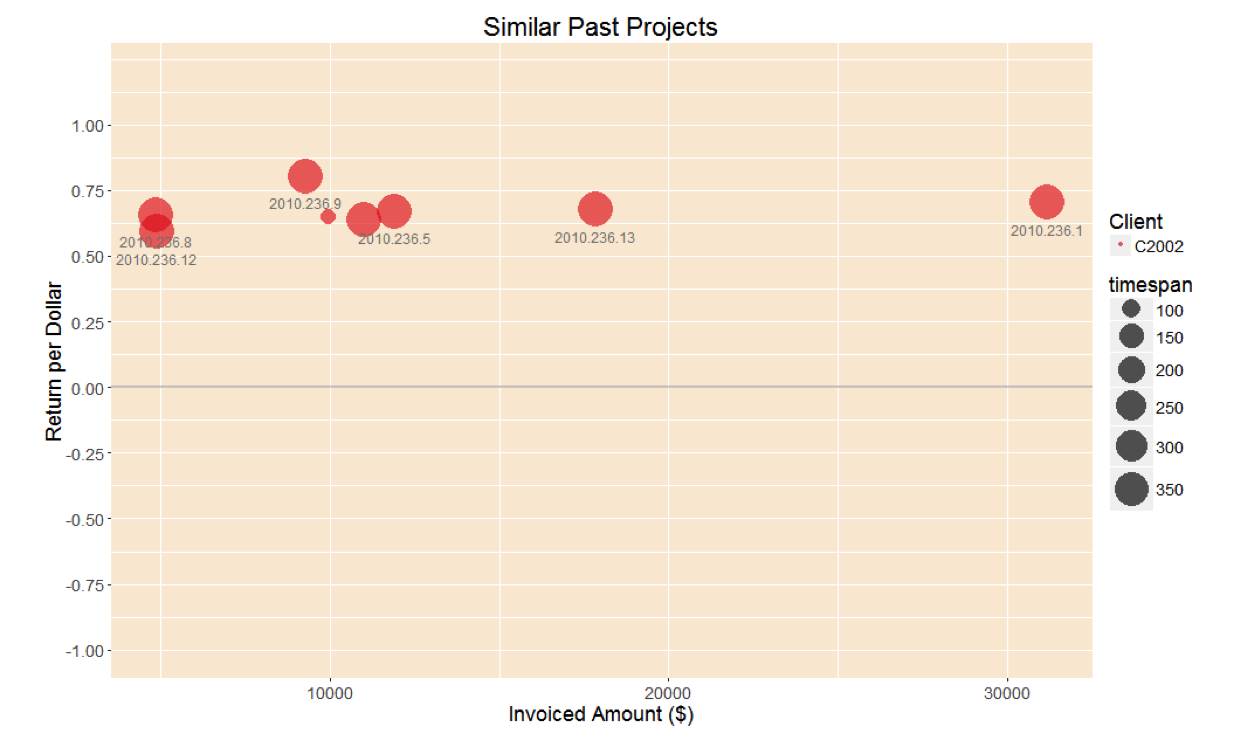
The improvements in profit sound promising, however it is important to consider possible limitations in the data which is discussed in the next section. One of the unknown factors in translating this research into business practises is how well users can guess the project variables. The data used to develop the model was obtained from the company's records of how projects played out and there are no records of initially expected values of variables. For example, if a project lasted 2 years and it was initially expected it would only take 1.5 years, the only data the company retains is that the project lasted two years. This begs the question, how well could a user predict the variables before beginning a project? Uncertainty has been managed to a degree by binning variables such as 'total amount invoiced' and 'timespan' to 5 or 6 broad categories (i.e. "1 day - 3 weeks", "3 weeks - 2.5 months", etc). However, two remaining categories: team size and percent of project completed by professional level employees were left as continuous variables. It would be interesting to test how well business managers could guess these four variables at the start of a project.

The process of guessing varibles at the beginning of a project and then waiting for the project to complete is a long checking process. However, it would be worthwhile integrating this data into the system's pre-existing data collecting processes. When a new project is entered into the database, it could be required that the project manager guesses the time-span (1 of 6 categories), total amount invoiced (1 of 5 categories), team size and percent to be completed by professionals. This data itself could feed into a revised version of the prediction algorithm.

Another limitation for these types of algorithms in general, as identified in the literature, is user uptake of the models and a lack of trust in the output (Akintoye and Fitzgerald 2000). It was for this reason the case study only tested models that provided insight into variable relationships (ensemble trees and logistic regression) as opposed to neural networks and SVM's that operate as 'black box' predictors. It is envisioned that a future user interface for this algorithm would display variable relationships such as this partial dependency plot produced by a boosted tree model that displays the probability of losing money vs. the team size:



Running a nearest-neighbour algorithm concurrently with the probability algorithm could provide a cross check of the algorithm's results and engage the user's past experiences in similar projects. This could work by presenting a graphical summary of the 4 or 5 most-similar past projects to the new project being assessed by the manager. Memories of how these similar projects played out would be triggered as well as communication with other manager's that led the similar projects. This would allow the user to connect specific past experience with the probability model output. Further development and testing is required to assess how much users trust the algorithm output and how trustworthiness may be improved.



# Conclusion

The aim of this project was to predict the profitability of projects for consulting businesses using their internal CRM data. Specifically, a statistical or machine learning model was created based on historical project data. The second aim was to test whether the model would have a positive impact on the overall profitability of a case study business. These research aims were fulfilled and the prediction model was refined to a binary classification problem of whether a project would be profitable or loss-making. This suited the limited data size best which was also highly complex. After testing several machine-learning and statistical models, a blended model of logistic regression, random forests, and boosted tree models predicted project profitability best. When applied to improving overall profitability of the business, the blended model which simply averaged the raw probability results from the three aforementioned models outperformed the individual models as well as the logistic regression blended model. Final results showed the model could improve profitability of the business by between 106.3% and 111.3% (the 95% confidence interval) by rejecting projects that have a 'probability' of being a loss-making job over 0.55 (approximately 4.2% of projects).

These results make a significant contribution to literature in this area as they provide a clear framework for how a business can realise tangible profit improvements using a set of algorithms and project data that is passively collected by businesses. Previous studies verified algorithms' prediction accuracy in cost estimation but stopped short of how the algorithms could be implemented in every day decision making and the affect on budgets. Additionally, this research tested ensemble tree methods and furthermore blended models on the problem of cost estimation. Past research had largely experimented with regression models and to some extent neural networks and case based reasoning. The advantage of ensemble tree methods is that they provide insight into the model's reasoning while having comparable predictive power to neural networks. Having said that, a limitation to the study was that there was not sufficient data to utilise the strength of ensemble tree methods and logistic regression performed simiarly. However, the final highest performing model was a blend of regression and two ensemble tree methods.

There are several aspects of this work that should be investigated further including how well managers can estimate project input variables before a project is won. Particularly, estimations of the timespan, team size, total invoiced amount, and percentage of hours completed by professionals should be tested. These initial estimates could also be included in the each project's data entry for future analysis. Finally, research into user uptake of the model should be assessed. It is likely that in order to establish a manager's trust with the algorithm, additional model output will be required such as partial dependency plots from the boosted trees. Also, a 'nearest neighbour' algorithm could be run concurrently to show the manager outcomes of past similar projects, which should align with the model's probabiltiy prediction. Overall, this project has successfully demonstrated a blended statistical and machine learning model has the potential to aid manager's in rejecting loss-making projects to improve their business's profitability.

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](https://doi.org/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](https://doi.org/http://dx.doi.org/10.1016/j.ijproman.2010.04.003).

Bergeron, François, and Jean-Yves St-Arnaud. 1992. “Estimation of Information Systems Development Efforts: A Pilot Study.” Journal Article. *Information & Management* 22 (4): 239–54.

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

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

———. 2001b. “Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author).” Journal Article. *Statistical Science* 16 (3): 199–231.

Breiman, Leo, and Adele Cutler. 2005. “Random Forests.” Journal Article. <https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm>.

Byun, Hyeran, and Seong-Whan Lee. 2002. “Applications of Support Vector Machines for Pattern Recognition: A Survey.” Book Section. In *Pattern Recognition with Support Vector Machines*, 213–36. Springer.

Carrizosa, Emilio, Belen Martin-Barragan, and Dolores Romero Morales. 2010. “Binarized Support Vector Machines.” Journal Article. *INFORMS Journal on Computing* 22 (1): 154–67.

Caruana, Rich, and Alexandru Niculescu-Mizil. n.d. “An Empirical Comparison of Supervised Learning Algorithms.” Conference Proceedings. In, 148:161–68. ACM. doi:[10.1145/1143844.1143865](https://doi.org/10.1145/1143844.1143865).

Champely, Stephane. 2015. *Pwr: Basic Functions for Power Analysis*. <https://CRAN.R-project.org/package=pwr>.

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

Davenport, Thomas H, and Jeanne G Harris. 2007. *Competing on Analytics: The New Science of Winning*. Book. Harvard Business Press.

Dissanayaka, Sunil M, and Mohan M Kumaraswamy. 1999. “Evaluation of Factors Affecting Time and Cost Performance in Hong Kong Building Projects.” Journal Article. *Engineering Construction and Architectural Management* 6 (3): 287–98.

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

Flyvbjerg, Bent, Massimo Garbuio, and Dan Lovallo. 2014. “Better Forecasting for Large Capital Projects.” Journal Article. *McKinsey on Finance* 52: 7–13.

Harrell, Frank E. 2013. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Book. Springer Science & Business Media.

Harris, Hank. 1999. “Risky Business.” Journal Article. *Civil Engineering (American Society of Civil Engineers) [H.W. Wilson - AST]* 69 (1): 63. <http://qut.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwY2AwNtIz0EUrE5INLFKBlZNFoqllqmVySpJpkoVRYgqwakw2NzcwS0tCG3-DnQoBjW1YIQkuuVPyk0GD5sBeO7AtAzphyr6gUBd0ixRothV6pQYzAyuw4jQBnaVvHhQIL5ktzME3qQEzlqmuuQFokzta-QuuVNwEGGALCmBDXDkYewgwD20k162CDPzQtqeCIySxCDEwpeYJM3AjnUgowsAXlFmcXakAvZqyWJRBxs01xNlDF2ZffEpOTryxhbmlIWimTIyBNxG0Qj6vBLyTLkWCQSEtGViuJRqlAOv8ZJM00GyepbGBQZqpebKFeZq5UZIkgyhWs6RwiEszcEGOMAANR8gwsJQUlabKgrdLAgDi5pbP>.

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-BrzBiPSwRH8tocBGfgTxSlGFGrbpQQFAAGIV52JmKxhByhShS_lQg53OGzhsR9IHSKJ6YsfDxcttlP8vYzzrrTEJ4vjYNGaUN8EcoeYfpUofnMCyKACQZ-O6qMEelNAQVpUEpjR1VUGrnKIpucN_BcvnlzqK86ZUNR4cBwrbpes4tFjlf4Yfds9R8fYH9dAA8FGnufK71MnR_ArkqHO7iTW5RQEm_uU4aMEPxsA9subLFFGbHpMYwfaRJ9ljaIkeqaQYt1tAWqU91odzNCXkH0KgCja453QKNKtDuKEJGJWRdqgHrUoCLlnB1qQSLFmCdkpuH8Ww4MXemu_iUBUYW8r_YZ6QRYppDmot0yMQgNQ6myQzkCQZ8l0EO3oKnkT95HMphUw17G5Gz1_vKDaAlwrJNt-edE2pb3PEiL-aRkzh-AKfFqO8nPse2BbFt9S9I-69JXf59-4oclpbTJvt59s2uRQ7tL6t_R8w>.

Haykin, Simon, and Neural Network. 2004. “A Comprehensive Foundation.” Journal Article. *Neural Networks* 2 (2004).

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

Heemstra, Fred J. 1992. “Software Cost Estimation.” Journal Article. *Information and Software Technology* 34 (10): 627–39.

Hothorn, Torsten, Kurt Hornik, and Achim Zeileis. 2006. “Unbiased Recursive Partitioning: A Conditional Inference Framework.” Journal Article. *Journal of Computational and Graphical Statistics* 15 (3): 651–74.

Jed Wing, Max Kuhn. Contributions from, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, et al. 2015. *Caret: Classification and Regression Training*. <https://CRAN.R-project.org/package=caret>.

Jorgensen, M., Magne Jorgensen, M. Shepperd, and Martin Shepperd. 2007. “A Systematic Review of Software Development Cost Estimation Studies.” Journal Article. *IEEE Transactions on Software Engineering* 32; 33 (1): 33–53. doi:[10.1109/TSE.2007.256943](https://doi.org/10.1109/TSE.2007.256943).

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](https://doi.org/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-GnfNLuKvBChkzFTJcgnkVadGg1XkVllKiXYEDHYLYxKUK3imPCfGklIwA8KxYJQpSsAaSHqIHmmXFLGpffJkEaG9EhayjRyriOBfRmj_izpfyNOzeXt50F9OK5_hN_1ZRzC7fh_EfCpeICxEIg2fcebqtoHtzlOic26UYobKWZFMUOQmPXObor7RJiMJT0BSJGqCDnscsmK1yqgrlQfklMBXWlSyH209lQwoEgPdDnfetDCFO2lWpRmoVBDJqRBOq9a3NTxgZ1i_cOABY3jDfSdkYIhgzEueCSL_MmzeFcl3xSHql3f-6BF6OGyaV-h-fdPY1z5t-DdLboxJ>.

Love, Peter ED, YC Raymond, and David J Edwards. 2005. “Time–cost Relationships in Australian Building Construction Projects.” Journal Article. *Journal of Construction Engineering and Management*.

Lunney, Gerald H. 1970. “Using Analysis of Variance with a Dichotomous Dependent Variable: An Empirical Study1.” Journal Article. *Journal of Educational Measurement* 7 (4): 263–69.

Ma, Zhiliang, and Zhe Liu. 2014. “BIM-Based Intelligent Acquisition of Construction Information for Cost Estimation of Building Projects.” Journal Article. *Procedia Engineering* 85: 358–67.

Markham, Selby. n.d. “Analysis of Variance (ANOVA).” Web Page. Monash University. <http://www.csse.monash.edu.au/~smarkham/resources/anova.htm>.

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. n.d. “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.

Moløkken, Kjetil, and Magne Jørgensen. n.d. “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/eLvHCXMwnV1Nb9QwEB0VekFClPLVAJV8gGPAH4kbc4GlyraIbhftpurRysTe06os3VTqz8fjTbawBanqKZHiOIntvBmPPe8BKPmBpxuYgMHsmgALqtHauDrYJRTaZNlB4b3XiBvxty99pGDV2z1IRuR2PxsKmn8UAUm1IcKwz4tfKclI0XJrp6nxALaJVoSG_Ej-WC8rFHnUE4x_Fmlv_e1WUh7I1fKfFiham-EO9Nuf-tjX_FZywW02x3t_xFN40nmlbLAaRruw5S-ewU6v-MA6AHgOn6bjYXU-mJQs-LfjScXKafVtFMNcjAQ8jtiAnZZnk8FJOFTn48l3Vp5Oy9HXk_IFnA3L6vA47dQX0gWlFaWNwjwjOjfDZ43hssFaovNCu7wIZwpnJrhmuSPtK2xE0QhZ68IJn3PnMzxQL-FxTbv0L9qYzef2gKkw-eJZ1BnFzGcKQ_Ua3SxUyRELTGC_byZbI8V6mnZpbxopgVfr624-t2G-QynmSooE3q-6zy5WLB1W2qW03BYBoSTdrbltr9sE9jbKkaChKLgyCbz7s-PXBWjaGJ6guInGPQFxl2KHHcU6UQu0r__73m_gkYwCG7QB-C08bC-v_H5MOv0NwrD2yw>.

Pinto, Jeffrey K, and Dennis P Slevin. n.d. “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>.

Putler, Daniel S, and Robert E Krider. 2012. “Customer and Business Analytics: Applied Data Mining for Business Decision Making Using R.” Online Multimedia. CRC Press.

Radenkovic, Predrag. 2010. “Random Forest.” Online Multimedia. University of Belgrade.

Rice, William R. 1989. “Analyzing Tables of Statistical Tests.” Journal Article. *Evolution* 43 (1): 223–25. <http://qut.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwlV1LT8MwDLbQJCQuPIcYY1IvHAvNo0l6Qggx8QPGOcqTy7SNbRzg1-OkrZjYAfEDUrluY3-2P9sAjN5V5S-bwGiD2Fc6g59fsWAqYZTxtYvC15zkAupu_q1f6phIlpklmGv6CJfsPNyjD8IYQz6s3su0PCoVWbtNGmiJE15uiX07o3dJC4IJL5miP8a3JSDuGeHsWaYn0JNE-zzXfK-RYH9y4z8FPoXjDncWj-2PcgYHYXEOh-0mys8LuM3DSb7QjxWz9LRNsYxFAqJ5jjMenKHgmyG8Tp9nTy9lt0GhfCMELZmPqQ0nChpS3GW5iZbLKkTPee2slCYahpCHWhV45D6EJtVBhTVWEWNZwy5hsFguwhUUSnpHJCPBu4Y3xCkaiRHB11EZpQIdwTBpVad7sV0bp_GVBSFiBJOsEL1qp2dojDoS-Vx3WhnBuFe_7q7PRlMqKxSxqq__OD2Go5b0lXIiNzDYrj_CJPdsfgOglrtM>.

Robin, Xavier, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. 2011. “PROC: An Open-Source Package for R and S+ to Analyze and Compare ROC Curves.” *BMC Bioinformatics* 12: 77.

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>.

Seng, Jia-Lang, and TC Chen. 2010. “An Analytic Approach to Select Data Mining for Business Decision.” Journal Article. *Expert Systems with Applications* 37 (12): 8042–57.

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. n.d. “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](https://doi.org/10.1155/2015/149702).

Strobl, Carolin, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. 2007. “Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution.” Journal Article. *BMC Bioinformatics* 8 (1): 1.

Strobl, Carolin, Torsten Hothorn, and Achim Zeileis. 2009. “Party on!” Journal Article.

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.

Tsai, Chih-Fong, and Yen-Jiun Chiou. 2009. “Earnings Management Prediction: A Pilot Study of Combining Neural Networks and Decision Trees.” Journal Article. *Expert Systems with Applications* 36 (3): 7183–91.

Tukey, John W. 1977. *Exploratory Data Analysis*. Book. Reading, Mass: Addison-Wesley Pub. Co. <http://qut.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwjVy9DoIwEL4oLrioiBHFhBfAQAuknY3Exc2dAG1HF98_8UprDUYTx-vQv1yv9_d9AJQcs_TDJqgcjSXT7F5F0SrU6V52oisr3vZlb9Bkf-Tf3i38LoWBKkdLTsdIjilMK0odniPTFV8M1yvL8eTkEbxv-FDqJXgaZLCCibwHsLCuYGIf2iOA-dXRqaLka5fQMCqvITJtc0N1PNENnklriUVCiOvz7XRJcbXG5mWa195zsgEPY325hYQpSXTIJvG-CnQNGGeC68piRgQRTEYQfp1i92N8D36OHp7JGcQwU6jT8jCc-gkw6HHX>.

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

Wu, Paul Pao-Yen, and Kerrie Mengersen. 2013. “A Review of Models and Model Usage Scenarios for an Airport Complex System.” Journal Article. *Transportation Research Part A: Policy and Practice* 47: 124–40.